Original Article

Comparison of Machine-Learning Algorithms for the Prediction of Current Procedural Terminology (CPT) Codes from Pathology Reports

Joshua Levy[1,2,3], Nishitha Vattikonda[4], Christian Haudenschild[5], Brock Christensen[2,6,7], Louis Vaickus[1]

[1]Emerging Diagnostic and Investigative Technologies, Clinical Genomics and Advanced Technologies, Department of Pathology and Laboratory Medicine, Dartmouth Hitchcock Medical Center, Lebanon, New Hampshire, USA, [2]Department of Epidemiology, Geisel School of Medicine at Dartmouth, Lebanon, New Hampshire, USA, [3]Program in Quantitative Biomedical Sciences, Geisel School of Medicine at Dartmouth, Lebanon, New Hampshire, USA, [4]Thomas Jefferson High School for Science and Technology, Alexandria, Virginia, USA, [5]University of Minnesota Medical School, Minneapolis, Minnesota, USA, [6]Department of Molecular and Systems Biology, Geisel School of Medicine at Dartmouth, Lebanon, New Hampshire, USA, [7]Department of Community and Family Medicine, Geisel School of Medicine at Dartmouth, Lebanon, New Hampshire, USA

Submitted: 29-July-2021 **Revised:** 20-November-2021 **Accepted:** 30-November-2021 **Published:** 05-January-2022

Abstract

Background: Pathology reports serve as an auditable trial of a patient's clinical narrative, containing text pertaining to diagnosis, prognosis, and specimen processing. Recent works have utilized natural language processing (NLP) pipelines, which include rule-based or machine-learning analytics, to uncover textual patterns that inform clinical endpoints and biomarker information. Although deep learning methods have come to the forefront of NLP, there have been limited comparisons with the performance of other machine-learning methods in extracting key insights for the prediction of medical procedure information, which is used to inform reimbursement for pathology departments. In addition, the utility of combining and ranking information from multiple report subfields as compared with exclusively using the diagnostic field for the prediction of Current Procedural Terminology (CPT) codes and signing pathologists remains unclear. **Methods:** After preprocessing pathology reports, we utilized advanced topic modeling to identify topics that characterize a cohort of 93,039 pathology reports at the Dartmouth-Hitchcock Department of Pathology and Laboratory Medicine (DPLM). We separately compared XGBoost, SVM, and BERT (Bidirectional Encoder Representation from Transformers) methodologies for the prediction of primary CPT codes (CPT 88302, 88304, 88305, 88307, 88309) as well as 38 ancillary CPT codes, using both the diagnostic text alone and text from all subfields. We performed similar analyses for characterizing text from a group of the 20 pathologists with the most pathology report sign-outs. Finally, we uncovered important report subcomponents by using model explanation techniques. **Results:** We identified 20 topics that pertained to diagnostic and procedural information. Operating on diagnostic text alone, BERT outperformed XGBoost for the prediction of primary CPT codes. When utilizing all report subfields, XGBoost outperformed BERT for the prediction of primary CPT codes. Utilizing additional subfields of the pathology report increased prediction accuracy across ancillary CPT codes, and performance gains for using additional report subfields were high for the XGBoost model for primary CPT codes. Misclassifications of CPT codes were between codes of a similar complexity, and misclassifications between pathologists were subspecialty related. **Conclusions:** Our approach generated CPT code predictions with an accuracy that was higher than previously reported. Although diagnostic text is an important source of information, additional insights may be extracted from other report subfields. Although BERT approaches performed comparably to the

Address for correspondence: Dr. Joshua Levy, Emerging Diagnostic and Investigative Technologies, Clinical Genomics and Advanced Technologies, Department of Pathology and Laboratory Medicine, Dartmouth Hitchcock Medical Center, 1 Medical Center Drive, Borwell Building 4th Floor, Lebanon, NH 03766, USA.
E-mail: joshua.j.levy@dartmouth.edu

How to cite this article: Levy J, Vattikonda N, Haudenschild C, Christensen B, Vaickus L. Comparison of machine-learning algorithms for the prediction of current procedural terminology (CPT) codes from pathology reports. J Pathol Inform 2022;13:3.
Available FREE in open access from: http://www.jpathinformatics.org/text.asp?2022/13/1/3/334788

Access this article online

Quick Response Code:

Website: www.jpathinformatics.org

DOI: 10.4103/jpi.jpi_52_21

XGBoost approaches, they may lend valuable information to pipelines that combine image, text, and -omics information. Future resource-saving opportunities exist to help hospitals detect mis-billing, standardize report text, and estimate productivity metrics that pertain to pathologist compensation (RVUs).

Keywords: BERT, current procedural terminology, deep learning, machine learning, pathology reports, XGBoost

Background and Significance

Electronic Health Records (EHR)[1] refer to both the structured and unstructured components of patients' health records/information (PHI), synthesized from a myriad of data sources and modalities. Such data, particularly clinical text reports, are increasingly relevant to "Big Data" in the biomedical domain. Structured components of EHR, such as clinical procedural and diagnostic codes, are able to effectively store the patient's history,[2-4] whereas unstructured clinical notes reflect an amalgamation of more nuanced clinical narratives. Such documentation may serve to refresh the clinician on the patient's history, highlight key aspects of the patient's health, and facilitate patient handoff among providers. Further, analysis of clinical free text may reveal physician bias or inform an audit trail of the patient's clinical outcomes for purposes of quality improvement. As such, utilizing sophisticated algorithmic techniques to assess text data in pathology reports may improve decision making and hospital processes/efficiency, possibly saving hospital resources while prioritizing patient health.

NLP[3,5-8] is an analytic technique that is used to extract semantic and syntactic information from textual data. Traditionally, rule-based approaches cross-reference and tabulate domain-specific key words or phrases with large biomedical ontologies and standardized vocabularies, such as the Unified Medical Language System (UMLS).[9,10] However, although these approaches provide an accurate means of assessing a narrow range of specified patterns, they are neither flexible nor generalizable since they require extensive annotation and development from a specialist. Machine-learning approaches (e.g. support vector machine (SVM), random forest)[11,12] employ a set of computational heuristics to circumvent manual specification of search criteria to reveal patterns and trends in the data. Bag-of-word approaches[13,14] study the frequency counts of words (unigrams) and phrases (bigrams, etc.) to compare the content of multiple documents for recurrent themes, whereas deep learning approaches[15-17] simultaneously capture syntax and semantics with artificial neural network (ANN) techniques. Recent deep learning NLP approaches have demonstrated the ability to capture meaningful nuances that are lost in frequency-based approaches; for instance, these approaches can effectively contextualize short- and long-range dependencies between words.[18,19] Despite potential advantages conferred from less structured approaches, the analysis of text across any domain usually necessitates balancing domain-specific customization (e.g. a medical term/abbreviation corpora) with generalized NLP techniques.

The analysis of pathology reports using NLP has been particularly impactful in recent years, particularly in the areas of information extraction, summarization, and categorization. Noteworthy developments include information extraction pipelines that utilize regular expressions (regex), to highlight key report findings (e.g., extraction of molecular test results),[20-23] as well as topic modeling approaches that summarize a document corpus by common themes and wording.[24] In addition to extraction methods, machine-learning techniques have been applied to classify pathologist reports[25]; notable examples include the prediction of ICD-O morphological diagnostic codes[26,27] and the prediction of CPT codes based only on diagnostic text.[28,29] Widespread misspelling of words and jargon specific to individual physicians have made it difficult to reliably utilize the rule-based and even machine-learning approaches for report prediction in a clinical workflow. In addition, hedging and uncertainty in text reports may further obfuscate findings.[30]

The CPT codes are assigned to report reimbursable medical procedures for diagnosis, surgery, and ordering of additional ancillary tests.[31,32] Assignments of CPT codes are informed by guidelines and are typically integrated into the Pathology Information System. As such, the degree to which new technologies and practices are implemented and disseminated are often informed by their impact on CPT coding practices. Reimbursements from CPT codes can represent tens to hundreds of millions of dollars of revenue at mid-sized medical centers, and thus systematic underbilling of codes could lead to lost hospital revenue, whereas overbilling patterns may lead to the identification of areas of redundant or unnecessary testing (e.g., duplication of codes, ordering of unnecessary tests, or assignment of codes representing more complex cases, etc.).

Ancillary CPT codes represent procedural codes that are automatically assigned when ancillary tests are ordered (e.g., immunohistochemical stains; e.g., CPT 88341, 88342, 88313, 88360, etc.). In contrast, primary CPT codes (e.g., CPT 88300, 88302, 88304, 88305, 88307, and 88309) are assigned based on the pathologist examination of the specimen, where CPT 88300 represents an examination without requiring the use of a microscope (gross examination), whereas CPT 88302–88309 include gross and microscopic examination of the specimen and are ordered by the case's complexity level (as specified by the CPT codebook; an ordinal outcome; e.g., *CPT 88305: Pathology examination of tissue using a microscope, intermediate complexity*), which determines reimbursement. The assignment of such codes is not devoid of controversy. Although it is

expected that raters will not report a specimen with a higher/lower code level, some may argue that such measures may not reflect the degree of difficulty for a particular case or there may not be a specific language that denotes primary CPT code placement of the phenomena (i.e., unlisted specimen, where it is at the pathologist's discretion to determine placement). For these codes, case complexity may ultimately be traced back to the clinical narrative reported in the pathology report text.[33]

Since the assignment of case complexity is sometimes unclear to the practicing pathologist as guidelines evolve, the prediction of these CPT codes from the diagnostic text using NLP algorithms can be used to inform whether a code was assigned that matches the case complexity. Recently developed approaches to predict CPT codes demonstrate remarkable performance; however, they only rely on the first 100 words from the report text, do not compare across multiple state-of-the-art NLP prediction algorithms, and do not consider report text outside of the diagnosis section.[28] Further, report lexicon is hardly standardized, as it may be littered with language and jargon that is specific to the sign-out pathologist and may vary widely in length for the same diagnosis, which can make it difficult to build an objective understanding of the report text.

Comparisons of different algorithmic techniques and relevant reporting text to use for the prediction of primary CPT codes are essential to further understand their utility for curbing under/overbilling issues. In addition, contextualizing primary code findings by ancillary findings and building a greater understanding of how pathologists differ in their lexical patterns may provide further motivation for the standardization of reporting practices and how report text can optimize the ordering of ancillary tests.[34]

Objective

The primary objective of this study is to compare the capacity to delineate primary CPT procedural codes (CPT 88302, 88304, 88305, 88307, 88309) corresponding to case complexity across state-of-the-art machine-learning models over a large corpus of more than 93,039 pathology reports from the Dartmouth-Hitchcock Department of Pathology and Laboratory Medicine (DPLM), a mid-sized academic medical center. Using XGBoost, SVM, and BERT techniques, we hope to gain a better understanding of which algorithms are useful for predicting primary CPT codes representing case complexity, which will prove helpful for the detection of under/overbilling.

Secondary Objectives

We have formulated various secondary objectives that are focused on capturing additional components of reporting variation:

1. **Expanded reporting subfields:** Exploration of methods that incorporate other document subfields outside of the diagnostic text into the modeling approaches, which may contain additional information.
2. **Ancillary Testing Codes:** Predicting the assignment of 38 different CPT procedure codes, largely comprising secondary CPT codes, under the hypothesis that nondiagnostic text provides additional predictive accuracy as compared with primary CPT codes, which may rely more heavily on the diagnostic text. Although the prediction of whether an ancillary test was ordered via secondary CPT codes has limited potential for incorporation into the Pathology Information System, as these codes are automatically assigned after test ordering, prediction of the ancillary tests can provide an additional context for the prediction of primary codes.
3. **Pathologist-Specific Language:** Investigate whether the sign-out pathologist can be predicted based on word choice. Although the sign-out pathologist can be found through an SQL query in the Pathology Information System, we are interested in translating sign-outs to a unified language that is consistent across sign-outs (i.e., a similar lexicon across pathologists, given diagnosis, code assignments, and subspecialty). As an example, some pathologists may more verbosely describe a phenomenon that could be succinctly summarized to match a colleague's description, though this could be difficult to disentangle without a quantitative understanding of lexical differences. To do this, we need to identify several components of variation (i.e., within a subspecialty, where reports from pathologists may vary widely); we want to further understand this heterogeneity to standardize communications within our department.

Although the final two objectives (ancillary testing and pathologist prediction) can be resolved by using an SQL query, we emphasize that these secondary objectives were selected to better identify the potential sources of reporting inconsistency with the aim of informing optimal reporting standards rather than imputing information that can be readily queried through the Pathology Information System.

Approach and Procedure

Data acquisition

We obtained Institutional Review Board approval and accessed more than 96,418 pathologist reports from DPLM, collected between June 2015 and June 2020. We removed a total of 3,379 reports that did not contain any diagnostic text associated with CPT codes, retaining 93,039 reports (Supplementary Table 1). Each report was appended with metadata, including corresponding EPIC (EPIC systems, Verona, WI),[35] Charge Description Master (CDM), and CPT procedural codes, the sign-out pathologist, the amount of time to sign out the document, and other details. Fuzzy string matching using the *fuzzywuzzy* package was used to identify whether any pathologists' names were misspelled (or resolve potential last name changes) between documents.[36] First, all unique pathologist names were identified. Then, for each pair of names, the token sort ratio was calculated,

thresholded by whether the ratio exceeded 0.7 to establish a unipartite graph of pathologist names connected to their candidate duplicates. Finally, clusters of similar names were identified by using connected component analysis. In most cases, unique names were assigned to each cluster of names, though in select cases, names were kept separate.[37] The documents were deidentified by stripping all PHI-containing fields and numerals from the text and replacing with holder characters (e.g. 87560 becomes #####). As a final check, we used regular expressions (regex) to remove mentions of patient names in the report text. This was accomplished by first compiling and storing several publicly available databases of 552,428 first and last names (Supplementary Materials, section "Additional Information on Deidentification Approach"). Then, using regex, we searched for the presence of each first and last name in the report subsections and replaced names at matched positions with white spaces. However, we did not remove mention of the physicians and consulting pathologist. The information on the physicians and consulting pathologist were identified in the "ordered by," "reports to," and "verified by" fields of the pathology report using known personal identifiers. The deidentification protocol was approved by the Institutional Review Board, Office of Research Operations and Data Governance. A total of 17,744 first and last names were stripped from the in-house data.

Preprocessing

We used regular expressions (regex) to remove punctuation from the text, and the text was preprocessed by using the Spacy package,[38] to tokenize the text. We utilized Spacy's en_core_web_sm processing pipeline (https://spacy.io/models/en#en_core_web_sm) to remove English stop words and words shorter than three characters. Out of concern for removing pathologist lexicon germane to pathologist sign-out, for this preliminary assessment, we did not attempt to prune additional words from the corpus outside of the methods used to generate word frequencies for the bag of words approaches. We also split up each pathology report into their structured sections: Diagnosis, Clinical Information, Specimen Processing, Discussion, Additional Studies, Results, and Interpretation. This allowed for an equal comparison between the machine-learning algorithms. The deep learning algorithm BERT can only operate on 512 words at a time due to computational constraints (See the "Limitations" and Supplementary Materials section "Additional Information on BERT Pretraining"). Sometimes, the pathology reports exceeded this length when considering the entire document (1.77% exceeded 512 words) and as such these reports were limited to the diagnosis section (0.02% exceeded 512 words) when training a new BERT model (Supplementary Table 1; Supplementary Figure 1). We removed all pathology reports that did not contain a diagnosis section.

Characterization of the text corpus

After preprocessing, we encoded each report tabulating the occurrence of all contiguous one- to two-word sequences (unigram and bigrams) to form sparse count matrices, where each column represents a word or phrase and each row represents the document, and the value is the frequency of occurrence in the document. Although the term "frequency" may be representative of the distribution of words/phrases in a corpus, high-frequency words that are featured across most of the document corpus are less likely to yield an informative lexicon that is specific to a subset of the documents. To account for less important but ubiquitous words, we transformed raw word frequencies to term frequency inverse document frequency (tf-idf) values, which up-weights the importance of the word based on its occurrence within a specific document (term frequency), but down-weights the importance if the word is featured across the corpus (inverse document frequency) (see the Supplementary Material section "Additional Description of Topic Modeling and Report Characterization Techniques"). We summed the tf-idf value of each word across the documents to capture the word's overall importance across the reports and utilized a word cloud algorithm to display the relative importance of the top words.

After constructing count matrices, we sought to characterize and cluster pathology documents as they relate to each other and ascribe themes to the clusters. Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP)[39] dimensionality reduction was used to project the higher dimensional word frequency data into lower dimensions while preserving important functional relationships. Each document could then be represented by a 3D point in the Cartesian coordinate system; these points were clustered by using a density-based clustering algorithm called HDBSCAN[40] to simultaneously estimate characteristic groupings of documents while filtering out noisy documents that did not explicitly fit in these larger clusters. To understand which topics were generally present in each cluster, we deployed Latent Dirichlet Allocation (LDA),[13] which identifies topics characterized by a set of words, and then derives the distribution of topics over all clusters. This is accomplished via a generative model that attempts to recapitulate the original count matrix, which is further outlined in greater detail in the Supplementary Material section "Additional Description of Topic Modeling and Report Characterization Techniques." The individual topics estimated using LDA may be conceptualized as a Dirichlet/multinomial distribution ("weight" per each word/phrase) over all unigrams and bigrams, where a higher weight indicates membership in the topic. The characteristic words pertaining to each topic were visualized by using a word cloud algorithm. Finally, we correlated the CPT codes with clusters, topics, and select pathologists by using Point-Biserial and Spearman correlation measures[41] to further characterize the overall cohort.

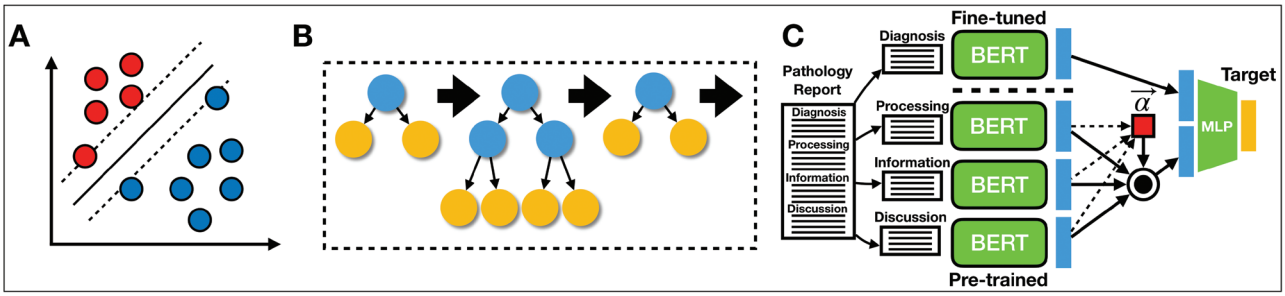


Figure 1: Model Descriptions: Graphics depicting: (A) SVM, where hyperplane linearly separates pathology reports, which are represented by individual datapoints; (B) XGBoost, which sequentially fits decision trees based on residuals from sum of conditional means of previous trees and outcomes; (C) *All-Fields* BERT model, where a diagnosis-specific neural network extracts relevant features from the diagnostic field, whereas a neural network trained on a separate clinical corpus extracts features for the remaining subfields; subfields are weighted and summed via the attention mechanism, indicated in red; subfields are combined with diagnostic features and fine-tuned with a multilayer perceptron for the final prediction

Machine learning models

We implemented the following three machine-learning algorithms in our study as a basis for our text classification pipeline [Figure 1]:

SVM

We trained an SVM[42,43] to make predictions by using the UMAP embeddings formed from the tf-idf matrix. The SVM operates by learning a hyperplane that obtains maximal distance (margin) to datapoints of a particular class [Figure 1A]. However, because datapoints/texts from different classes may not be separable in the original embedding space, the SVM model projects data to a higher dimensional space where data can be linearly separated. We utilized GPU resources via the ThunderSVM package[44] to train the model in reasonable compute time.

Bag of words with XGBoost

XGBoost algorithms[45] operate on the entire word by report count matrix and ensemble or average predictions across individual Classification and Regression Tree (CART) models.[46] Individual CART models devise splitting rules that partition instances of the pathology notes based on whether the count of a particular word or phrase in a pathology note exceeds an algorithmically derived threshold. Important words and thresholds (i.e. partition rules) are selected from the corpus based on their ability to partition the data, based on the purity of a decision leaf through the calculation of an entropy measure. Each successive splitting rule serves to further minimize the entropy or maximize the information gained. Random Forest models[47] bootstrap which subsets of predictors/words and samples are selected for a given splitting rule of individual trees and aggregate the predictions from many such trees; Extreme Gradient Boosting Trees (XGBoost) fit trees (structure and the conditional means of the terminal nodes) sequentially based on the residual (in the binary classification setting, misclassification is estimated using a Bernoulli likelihood) between the outcome and the sum of both the conditional means of the previous trees (which are set) and the conditional means of the current tree (which is optimized). This gradient-based optimization technique prioritizes samples with a large residual/gradient from the previous model fit to account for the previous "weak learners" [Figure 1B]. In both scenarios, random forest (a bagging technique) and XGBoost (a boosting technique), individual trees may exhibit bias but together cover a larger predictor space. Our XGBoost classifier models were trained by using the XGBoost library, which utilizes GPUs to speed up calculation.

BERT

ANN[48] are a class of algorithms that use highly interconnected computational nodes to capture relationships between predictors in complex data. The information is passed from the nodes of an input layer to the individual nodes of subsequent layers that capture additional interactions and nonlinearities between predictors while forming abstractions of the data in the form of intermediate embeddings. The BERT[18] model first maps each word in a sentence to its own embedding and positional vectors, which captures key semantic/syntactic and contextual information that is largely absent from the bag of words approaches. These word-level embeddings are passed to a series of self-attention layers (the Transformer component of the BERT model), which contextualizes the information of a single word in a sentence based on short- and long-term dependencies between all words from the sentence. The individual word embeddings are combined with the positional/contextual information, obtained via the self-attention mechanism, to create embeddings that represent the totality of a sentence. Finally, this information is passed to a series of fully connected layers that produce the final classification. With BERT, we are also able to analyze the relative importance and dependency between words in a document by extracting "attention matrices." We are also able to retrieve sentence-level embeddings encoded by the

network by extracting vectors from the intermediate layers before they pass for the final classification.

We trained the BERT models by using the *HuggingFace Transformers* package,[49] which utilizes GPU resources through the PyTorch framework. We used a collection of models that have already been pretrained on a large medical corpus[50] in order to both improve the predictive accuracy of our model and significantly reduce the computational load compared with training a model from scratch. Because significant compute resources are still required to train the model, most BERT models limit the document characterization length to 512 words. To address this, we split pathology reports into document subsections when training BERT models.

In training a BERT model, we updated the word embeddings through fine-tuning a pretrained model on our diagnostic corpus. This model, which had been trained solely on diagnostic text, could be used to predict the target of interest (*Dx Model*). However, we then used this fine-tuned model to extracted embeddings that were specific to the diagnosis subfield to serve as input for a model that could utilize text from other document subfields. We separately utilized the original pretrained model to extract embeddings from the other report subfields that are less biased by diagnostic codes and thus more likely to provide contextual information (*All Fields Model*). We developed a global/gating attention mechanism procedure that serves to dynamically prune unimportant, missing, or low-quality document subsections for classification [Figure 1C]. Predictions may be obtained when some/all report subfields are supplied via the following method:

$$y = f_{all-fields}(\vec{x}) = f_{MLP}\left(\left[\begin{array}{c}\vec{z}_{fine-tuned\ bert,dx},\\ \sum_{section}\alpha_{section}\vec{z}_{pretrained\ bert,section}\end{array}\right]\right)$$

$$\vec{\alpha} = softmax\left(\{f_{gate}\left(\overrightarrow{z_{section}}\right)\forall sections\}\right) \in [0,1], \vec{\alpha} = 1$$

$$f_{gate}\left(\overrightarrow{z_{section}}\right) = W_2 BatchNorm1d\left(ReLU\left(W_1\overrightarrow{z_{section}}\right)\right)$$

where $\vec{z}$ represents the embeddings extracted from the pretrained and fine-tuned BERT embeddings on respective report subsections, and $\vec{\alpha}$ is a vector of attention scores between 0 and 1 that dictates the importance of particular subsections. These attention scores are determined by using a separate gating neural network, f_{gate}, which maps $\vec{z}$, a 768-dimensional vector to a scalar for each document subsection through two projection matrices: W_1 a 768-dimension (dimensionality of BERT embeddings) by 100-dimensional matrix, and W_2 a 100-dimension (dimensionality of BERT embeddings) by 1-dimensional matrix that generates the attention scores. A softmax transformation is used to normalize the scores between zero and one across the subsections. Finally, f_{MLP} are a set of fully connected layers that operate on the concatenation between the BERT embeddings that were fine-tuned on the diagnosis-specific section and those extracted by using the pre-trained BERT model on the other document subfields, as weighted by using the gated attention mechanism (Supplementary Section "Additional Description of Explanation Techniques"). To train this model, we experimented with an ordinal loss function,[51] based off of the proportional odds cumulative link model specification, which respects the ordering of the primary CPT codes by case complexity, though ultimately, we opted for using a Cross-Entropy loss since ordinal loss functions are not currently configured for the other machine-learning methods (e.g., XGBoost).

Prediction of primary current procedural terminology codes

We developed machine-learning pipelines to delineate primary CPT codes requiring examination with a microscope (CPT 88302, 88304, 88305, 88307, 88309) using BERT, XGBoost, and SVM, with reports selected based on whether they contained only one of the five codes (where the primary codes were present in the following proportions: CPT 88302:0.67%, 88304:6.59%, 88305:85.97%, 88307:6.32%, and 88309:0.44%). The prevalence of most of the five codes did not change over time (Supplementary Figure 2; Supplementary Table 2). Given the characterization of the aforementioned deep learning framework, we utilized a BERT model that was pretrained first on a large corpus of biomedical research articles from PubMed, and then pretrained by using a medical corpus of free text notes from an intensive care unit (MIMIC3 database; Bio-ClinicalBERT; Supplementary Materials section "Additional Information on BERT Pretraining").[50,52,53] Finally, the model was fine-tuned on our DHMC pathology report corpus (to capture institution-specific idiosyncrasies) for the task of classifying particular CPT codes from diagnostic text. XGBoost was trained on the original count matrix, whereas SVM was trained on a 6-dimensional UMAP projection; a UMAP projection was utilized for computational considerations. The models were evaluated by using five-fold cross-validation as a means to compare the model performances. Internal to each fold is a validation set used for identifying optimal hyperparameters (supplementary section "Additional Information on Hyperparameter Scans") through performance statistics and a held-out test set. For each approach, we separately fit a model considering only the Diagnosis text (*Dx* Models) and all of the text (*All Fields* Models) to provide additional contextual information. We calculated the Area Under the Receiver Operating Curve (AUC-Score; considers sensitivity/specificity of the model at a variety of probability cutoffs; anything above a 0.5 AUC is better than random), F1-Score (which considers the tradeoff between

sensitivity and specificity) and macro-averaged these scores across the five CPT codes, which gives greater importance to rare codes. Since codes are also ordered by complexity (ordinal variable), we also report a confusion matrix, which tabulates the real versus predicted codes for each approach and measures both a spearman correlation coefficient and linear-weighted kappa between predicted and real CPT codes as a means to communicate how the model preserves the relative ordering of codes (i.e., if the model is incorrect, better to predict a code of a similar complexity).

Ancillary testing current procedural terminology codes and pathologist prediction tasks

To contextualize findings for primary codes, these machine-learning techniques were employed to predict each of 38 different CPT codes (38 codes remained after removing codes that occurred less than 150 times across all sign-outs) (e.g., if the prediction of primary codes relies on the diagnostic section, do secondary codes rely on other document sections more?). The primary code model predicted a categorical outcome, whereas ancillary testing models were configured in the multitarget setting, where each code represents a binary outcome. We compared cross-validated AUC statistics between and across the 38 codes to further explore the reasons that some codes yielded lower scores than others. We also compared different algorithms via the sensitivity/specificity reported via their Youden's index (the optimal tradeoff possible between sensitivity and specificity from the receiver operating curve), averaged across validation folds.

We similarly trained all models to recognize the texts of the 20 pathologists with the most sign-outs to see whether the models could reveal pathologist-specific text to inform future efforts to standardize text lexicon. We retained reports from the 20 pathologists with the most sign-outs, reducing our document corpus from 93,039 documents to 64,583 documents, and we utilized all three classification techniques to predict each sign-out pathologist simultaneously. The selected pathologists represented a variety of specialties. Choosing only the most prolific pathologists removed the potential for biased associations by a rare outcome in the multiclass setting.

Model interpretations

Finally, we used shapley additive explanations (SHAP; a model interpretation technique that estimates the contributions of predictors to the prediction through credit allocation)[54] to estimate which words were important for the classification of each of these codes, visualized by using a word cloud. For the BERT model, we utilized the Captum[55] framework to visualize backpropagation from the outcome to predictors/words via IntegratedGradients[56] and attention matrices. Additional extraction of attention weights also revealed not only which words and their relationships contributed to the prediction of the CPT code (i.e. self-attention denotes word-to-word relationships), but also which document subfields other than the diagnosis field were important for assignment of the procedure code (i.e. global/gating attention prunes document subfields by learning to ignore irrelevant information; the degree of pruning can be extracted during inference). Further description of these model interpretability techniques (SHAP, Integrated Gradients, Self-Attention / "word-to-word", Attention) may be found in the supplementary material (section "Additional Description of Explanation Techniques: SHAP, Integrated Gradients, Self-Attention, Attention Over Pathology Report Subfields"). Pathologist-specific word choice was extracted by using SHAP/Captum from the resulting model fit and visualized by using word clouds and attention matrices.

Results

Corpus preprocessing and Uniform Manifold Approximation and Projection for Dimension Reduction results

After initial filtering, we amassed a total of 93,039 pathology reports, which were broken into the following subsections: Diagnosis, Clinical Information, Specimen Processing, Discussion, Additional Studies, Results, and Interpretation. The median word length per document was 119 words (Interquartile Range; IQR=90). Very few reports contained subfields that exceeded the length acceptable by the BERT algorithm (2% of reports containing a *Results* section exceeded this threshold; Supplementary Table 1; Supplementary Figure 1).

Displayed first are word clouds of the top 25 words in only the diagnostic document subsection [Figure 2A] and across all document subsections [Figure 2B], with their size reflecting their tf-idf scores [Figure 2A and B]. As expected, the diagnostic-field cloud contains words that are pertinent to the main diagnosis, whereas the all-field cloud contains words that are more procedural, suggesting that other pathology document subfields yield distinct and specific clinical information that may lend complementary information versus analysis solely on diagnostic fields. We clustered and visualized the diagnostic subsection and also all document subsections after running UMAP, which yielded 8 and 15 distinct clusters, respectively [Figure 2C and D]. The number of words per report correlated poorly with the number of total procedural codes assigned (Spearman $r = 0.066, p < 0.01$). However, when these correlations were assessed within the HDBSCAN report clusters (subset to reports within a particular cluster for cluster-specific trends), 33% of the all-fields report clusters reported moderate correlations (Supplementary Table 3). Interestingly, one of the eight report clusters from the diagnostic fields experienced a moderate negative correlation with the number of codes assigned.

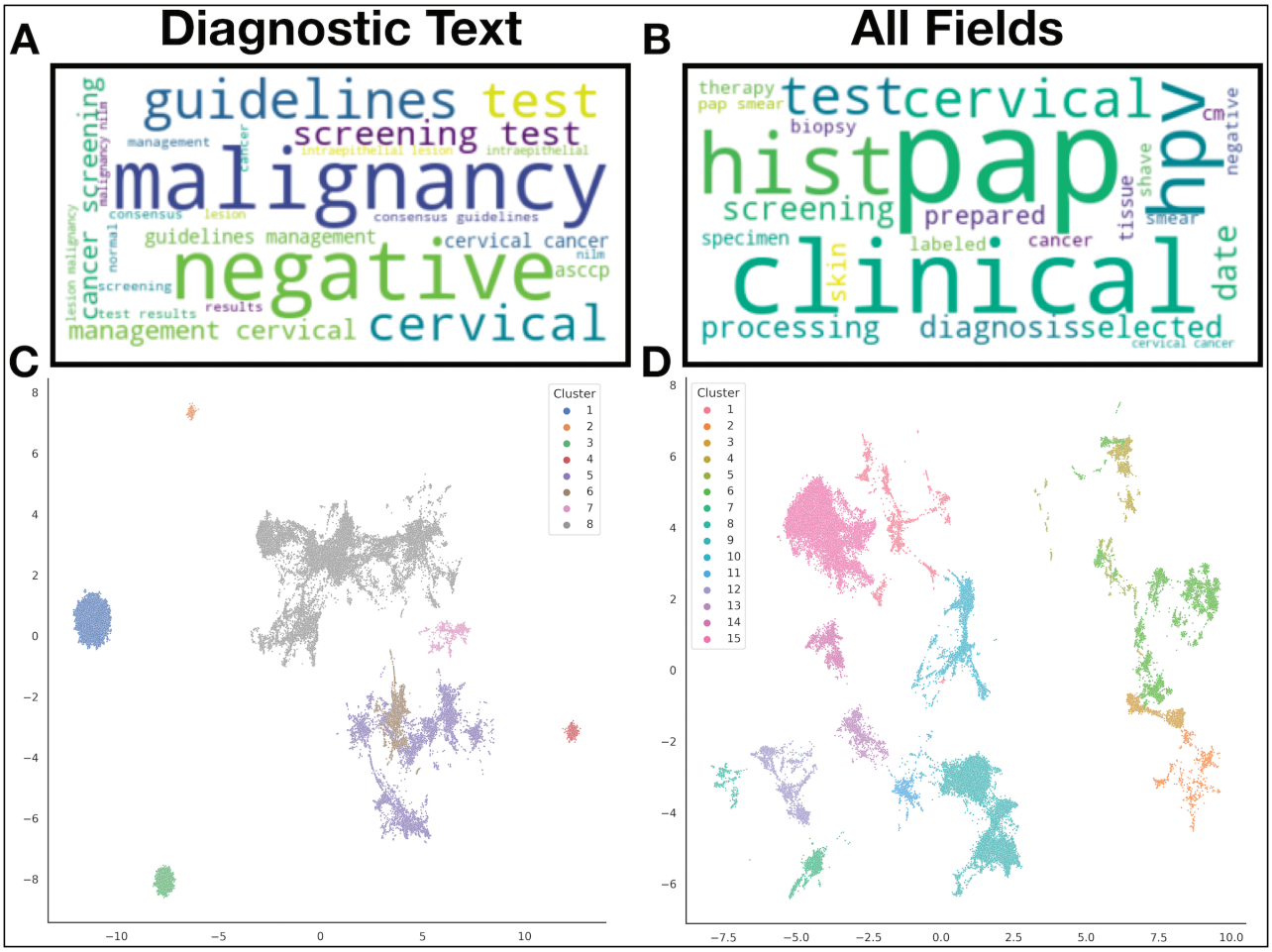


Figure 2: Pathology report corpus characterization: (A and B) Word cloud depicting words with the highest aggregated tf-idf scores across the corpus of: (A) diagnostic text only, (B) all report subfields (*all-fields*); important words across the corpus indicated by relative size of the word in the word cloud; (C and D) UMAP projection of the tf-idf matrix, clustered and noise removal via HDBSCAN for: (C) diagnostic texts only, and (D) all report subfields (*all-fields*)

Topic modeling with Latent Dirichlet Allocation and additional topic associations

From our LDA analysis on all document subsections, we discovered 10 topics [Figure 3; Supplementary Table 4]. Correlations between these topics with clusters, pathologists, and CPT codes are displayed in the supplementary material (Supplementary Figures 3–6). We discovered additional associations between CPT codes, clusters, and pathologists (Supplementary Figure 7A), suggesting a specialty bias in document characterization. We clustered pathologists using co-occurrence of procedural code assignments in order to establish "subspecialties" (e.g., pathologist who signs out multiple specialties) that could be used to help interpret sources of bias in an evaluation of downstream modeling approaches.

Primary current procedural terminology code classification results

The XGBoost and BERT models significantly outperformed the SVM model for the prediction of primary CPT codes [Table 1; Figure 4A and B; Supplementary Table 5]. The BERT model made more effective use of the diagnostic text (*macro-f1*=0.825; $\kappa = 0.852$) as compared with the XGBoost model (*macro-f1*=0.807; $\kappa = 0.835$). Incorporating the text from other report subfields provided only a marginal performance gain for BERT (*macro-f1*=0.829; $\kappa = 0.855$) and both a large and significant performance gain for XGBoost (*macro-f1*=0.831; $\kappa = 0.863$) [Figure 4A and B]. Across the BERT and XGBoost models, codes were likely to be misclassified if they were of a similar complexity [Table 1; Supplementary Table 5]. Plots of low-dimensional text embeddings extracted from the BERT *All-Fields* model demonstrated clustering by code complexity and relative preservation of the ordering of code complexity (i.e., reports pertaining to codes of lower/higher complexity clustered together) [Figure 4C].

Ancillary current procedural terminology code and pathologist classification results

We were able to accurately assign ancillary CPT codes to each document, regardless of which machine learning

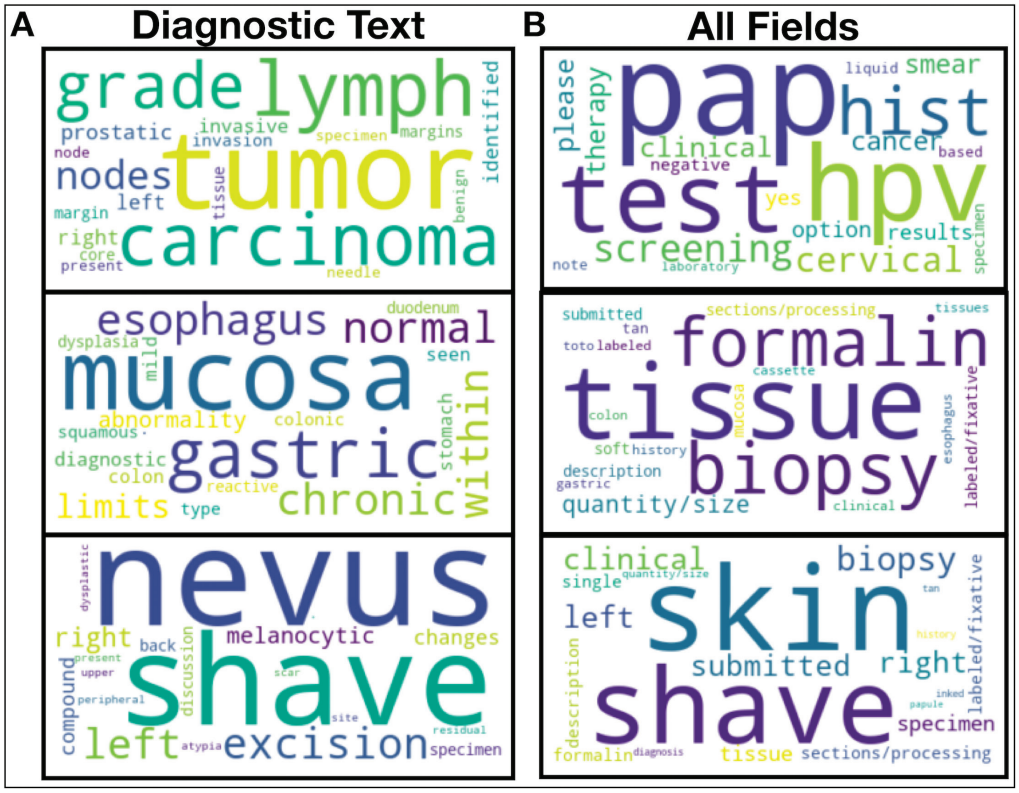


Figure 3: LDA Topic Words: Important words found for three select LDA Topics from: (A) diagnostic text only and (B) all report subfields (*all-fields*); important words across the corpus indicated by relative size of the word in the word cloud

Table 1: Predictive performances for primary CPT code algorithms

Approach	Type	Macro-F1 ± SE	κ ± SE	AUC ± SE	Spearman ± SE
BERT	Diagnosis	0.825 ± 0.0064	0.852 ± 0.0033	0.99 ± 0.0008	0.84 ± 0.0044
	All fields	0.828 ± 0.0062	0.855 ± 0.0032	0.99 ± 0.0006	0.843 ± 0.0044
XGBoost	Diagnosis	0.807 ± 0.0069	0.835 ± 0.0034	0.99 ± 0.0007	0.824 ± 0.0045
	All fields	0.832 ± 0.0069	0.863 ± 0.0032	0.994 ± 0.0004	0.855 ± 0.0042
SVM	Diagnosis	0.497 ± 0.0047	0.644 ± 0.0043	0.554 ± 0.0021	0.637 ± 0.0056
	All fields	0.518 ± 0.0048	0.668 ± 0.0044	0.554 ± 0.0014	0.652 ± 0.0058

Macro-F1 and AUC measures are agnostic to the ordering of the CPT code complexity; whereas Linear Kappa (κ) and Spearman correlation coefficients respect the CPT code ordering (88302, 88304, 88305, 88307, and 88309)

algorithm was utilized (Supplementary Figure 8; Supplementary Table 6). Across all ancillary codes, we found that XGBoost (median AUC=0.985) performed comparably to BERT (median AUC=0.990; P = 0.64) when predicting CPT codes based on the diagnostic subfield alone, whereas SVM performed worse (median AUC=0.966) than both approaches, per cross-validated AUC statistics (Supplementary Tables 6–10; Supplementary Figure 9). In contrast to results obtained for the primary codes, we discovered that classifying by including all of the report subelements (*All Fields*) performed better than just classifying based on the diagnostic subsection (P < 0.001 for both BERT and XGBoost approaches; Supplementary Tables 6, 8–10; Supplementary Figures 9 and 10), suggesting that these other more procedural / descriptive elements contribute meaningful contextual information for the assignment of ancillary CPT codes (Supplementary Materials section "Supplementary Ancillary CPT Code Prediction Results"). We also report that the sign-out pathologist can also be accurately identified from the report text, with comparable performance between the BERT (macro-f1=0.72) and XGBoost (macro-f1=0.71) models, and optimal performance when all report subfields are used (macro-f1=0.77 and 0.78, respectively) (Supplementary Materials section "Supplementary Pathologist Prediction Results"; Supplementary Table 11; Supplementary Figure 11).

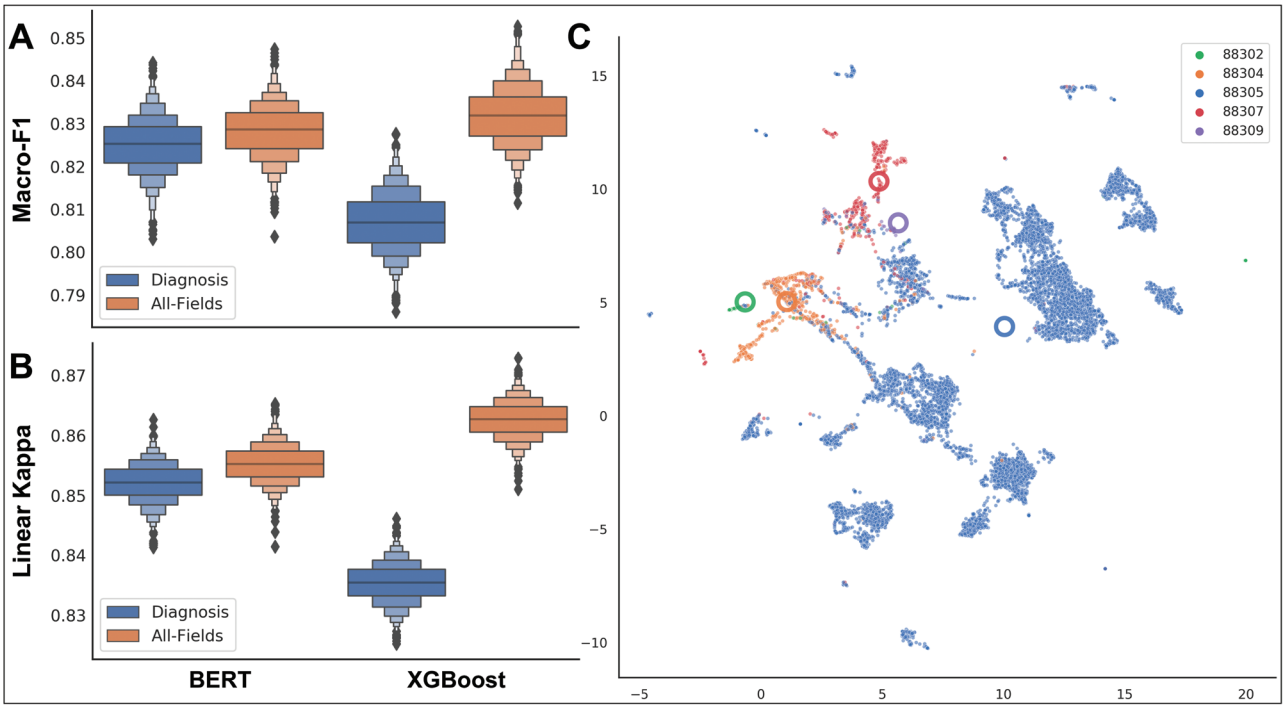


Figure 4: Primary CPT Code Model Performance: (A and B) Grouped boxenplots demonstrating the performance of machine-learning models (BERT, XGBoost) for the prediction of primary CPT codes (bootstrapped performance statistics; A) macro-averaged F1-Score, (B) Linear-Weighted Kappa for performance across different levels of complexity, which takes into account the ordinal nature of the outcome; reported across five CPT code), given analysis of either the diagnostic text (blue) or all report subfields (orange); (C) UMAP projection of *All-Fields* BERT embedding vectors after applying the attention mechanism across report subfields; each point is reported with information aggregated from all report subfields; individual points represent reports, colored by the CPT code; large thick circles represent the report centroids for each CPT code; note how codes CPT 88302 and CPT 88304 cluster together and separately CPT 88307 and CPT 88309 cluster together, whereas CPT 88305 sits in between clustered reports of low and high complexity

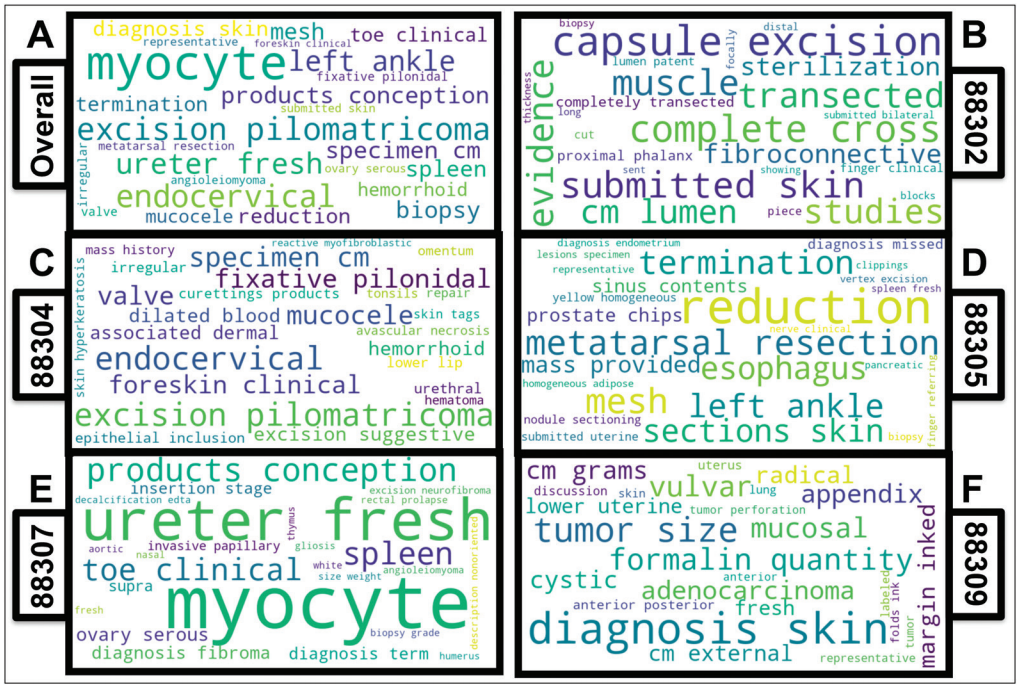


Figure 5: SHAP interpretation of XGBoost predictions: Word clouds demonstrating words found to be important using the XGBoost algorithm (*All-Fields*) for the prediction of primary CPT codes, found via shapley attribution; important words pertinent to each CPT code indicated by the relative size of the word in the word cloud; word clouds visualized for word importance (A) across all five primary CPT codes and (B–F) for the following CPT codes: (B) CPT code 88302; (C) CPT code 88304; (D) CPT code 88305; (E) CPT code 88307; and (F) CPT code 88309; note that the size of the word considers strength but not directionality of the relationship with the code, which may be negatively associated in some cases

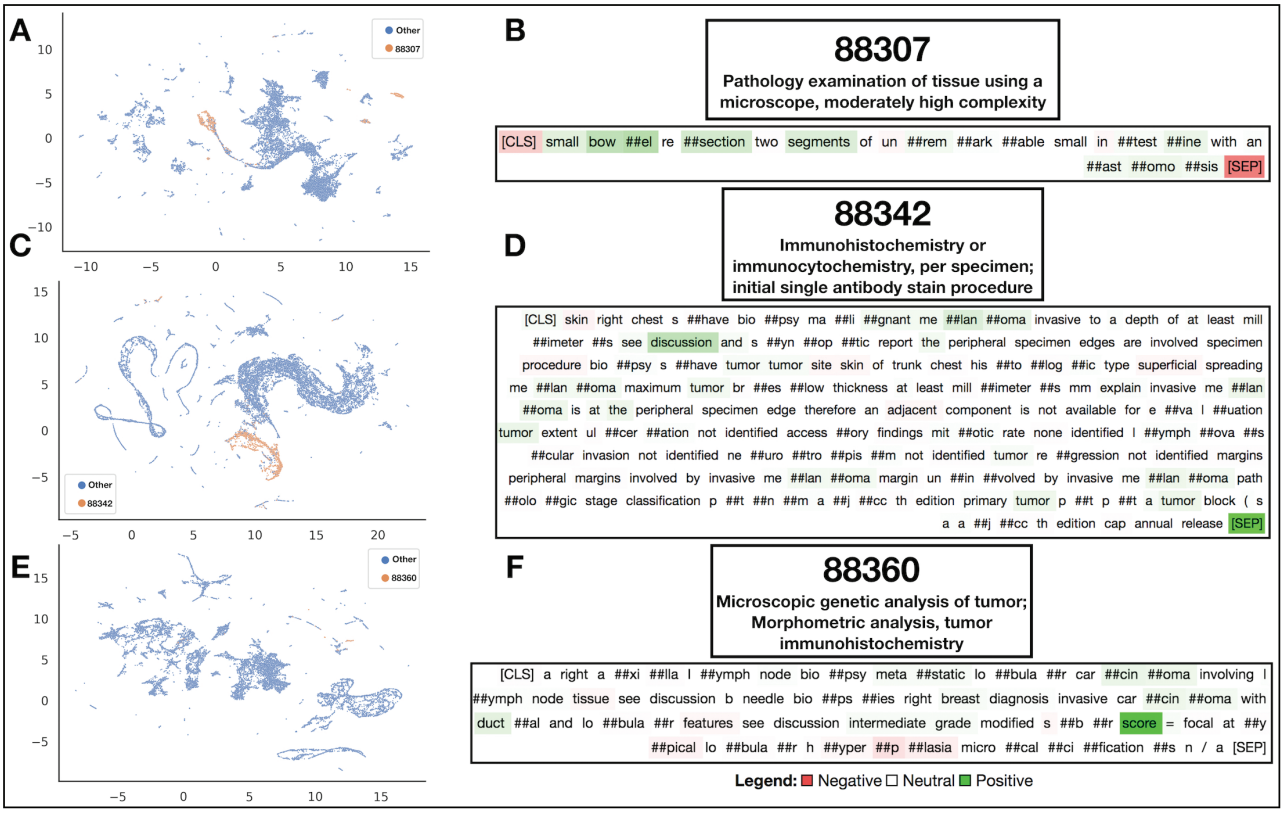


Figure 6: Embedding and Interpretation of BERT Predictions: (A, C, and E) UMAP projection of *All-Fields* BERT embedding vectors after applying the attention mechanism across report subfields; each point is reported with information aggregated from all report subfields; (B, D, and F) Select diagnostic text from individual reports interpreted by Integrated Gradients to elucidate words positively and negatively associated with calling the CPT code; Integrated Gradients was performed on the diagnostic text BERT models; Utilized CPT codes: (A and B) CPT code 88307, (C and D) CPT code 88342, and (E and F) CPT code 88360

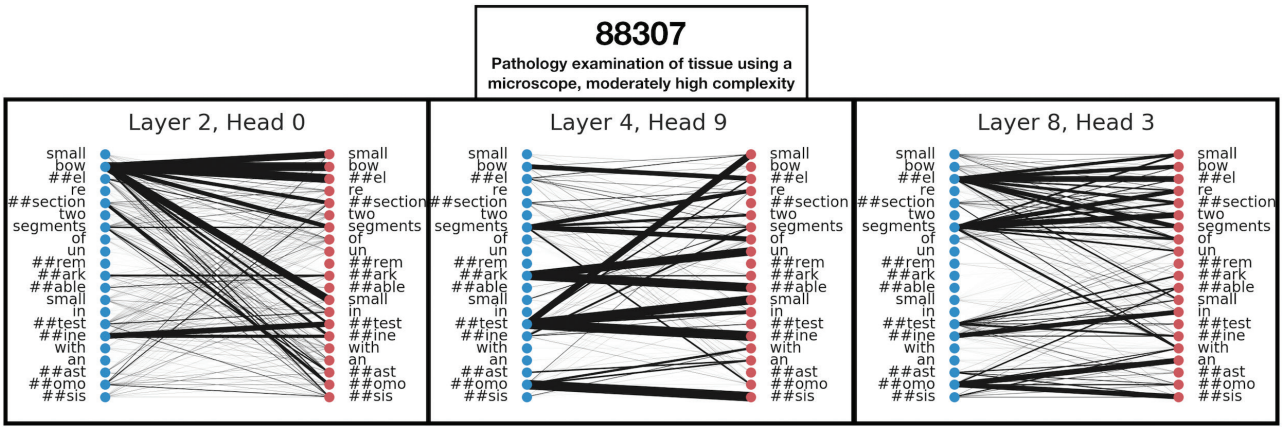
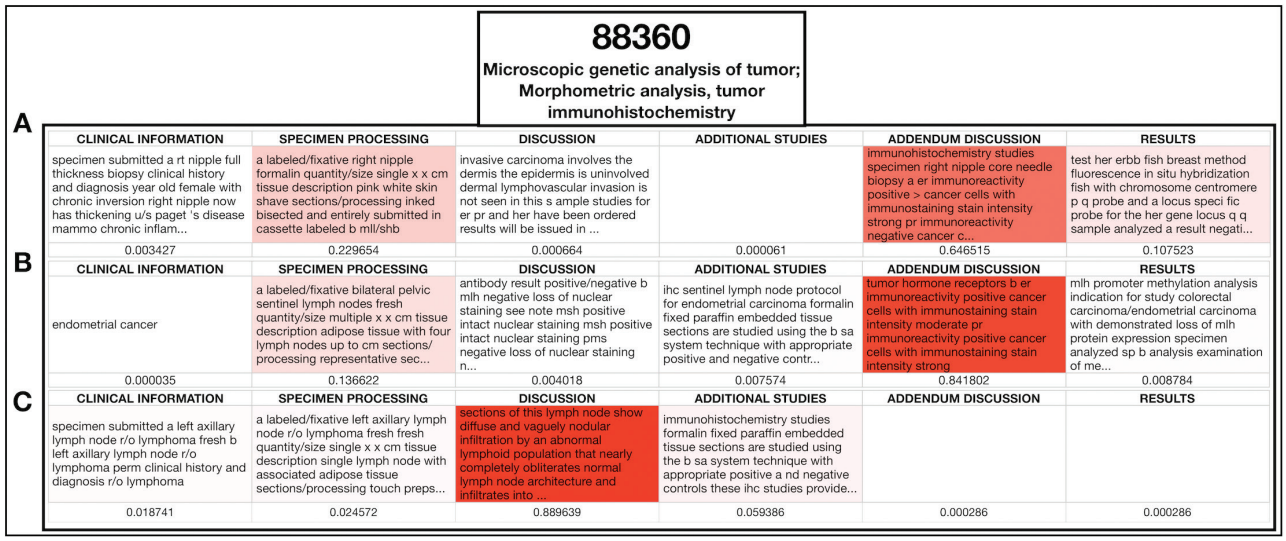


Figure 7: BERT Diagnostic Model Self-Attention: Output of self-attention maps for select self-attention heads/layers from the BERT diagnostic text model visualizes various layers of complex word-to-word relationships for the assessment of a select pathology report that was found to report CPT code 88307

Model interpretation results

We also visualized which words were found to be important for a subsample of primary and ancillary procedural codes by using the XGBoost algorithm [Figure 5; Supplementary Figure 12]. In the Supplementary Materials, we have also included a table that denotes the relevance of the top 30 words for the XGBoost *All Fields* model for the prediction of specific primary CPT codes, as assessed through SHAP (Supplementary Table 12). Reports that were assigned the same ancillary CPT code clustered together in select low-dimensional representations learned by some of the *All Fields* BERT models [Figure 6A, C, and E]. Model-based



88360

Microscopic genetic analysis of tumor; Morphometric analysis, tumor immunohistochemistry

A

CLINICAL INFORMATION	SPECIMEN PROCESSING	DISCUSSION	ADDITIONAL STUDIES	ADDENDUM DISCUSSION	RESULTS
specimen submitted a rt nipple full thickness biopsy clinical history and diagnosis year old female with chronic inversion right nipple now has thickening u/s paget 's disease mammo chronic inflam...	a labeled/fixative right nipple formalin quantity/size single x x cm tissue description pink white skin shave sections/processing inked bisected and entirely submitted in cassette labeled b mll/shb	invasive carcinoma involves the dermis the epidermis is uninvolved dermal lymphovascular invasion is not seen in this s ample studies for er pr and her have been ordered results will be issued in ...		immunohistochemistry studies specimen right nipple core needle biopsy a er immunoreactivity positive > cancer cells with immunostaining stain intensity strong pr immunoreactivity negative cancer c...	test her erbb fish breast method fluorescence in situ hybridization fish with chromosome centromere p q probe and a locus speci fic probe for the her gene locus q q sample analyzed a result negati...
0.003427	0.229654	0.000664	0.000061	0.646515	0.107523

B

CLINICAL INFORMATION	SPECIMEN PROCESSING	DISCUSSION	ADDITIONAL STUDIES	ADDENDUM DISCUSSION	RESULTS
endometrial cancer	a labeled/fixative bilateral pelvic sentinel lymph nodes fresh quantity/size multiple x x cm tissue description adipose tissue with four lymph nodes up to cm sections/ processing representative sec...	antibody result positive/negative b mlh negative loss of nuclear staining see note msh positive intact nuclear staining msh positive intact nuclear staining pms negative loss of nuclear staining n...	ihc sentinel lymph node protocol for endometrial carcinoma formalin fixed paraffin embedded tissue sections are studied using the b sa system technique with appropriate positive and negative contr...	tumor hormone receptors b er immunoreactivity positive cancer cells with immunostaining stain intensity moderate pr immunoreactivity positive cancer cells with immunostaining stain intensity strong	mlh promoter methylation analysis indication for study colorectal carcinoma/endometrial carcinoma with demonstrated loss of mlh protein expression specimen analyzed sp b analysis examination of me...
0.000035	0.136622	0.004018	0.007574	0.841802	0.008784

C

CLINICAL INFORMATION	SPECIMEN PROCESSING	DISCUSSION	ADDITIONAL STUDIES	ADDENDUM DISCUSSION	RESULTS
specimen submitted a left axillary lymph node r/o lymphoma fresh b left axillary lymph node r/o lymphoma perm clinical history and diagnosis r/o lymphoma	a labeled/fixative left axillary lymph node r/o lymphoma fresh fresh quantity/size single x x cm tissue description single lymph node with associated adipose tissue sections/processing touch preps...	sections of this lymph node show diffuse and vaguely nodular infiltration by an abnormal lymphoid population that nearly completely obliterates normal lymph node architecture and infiltrates into ...	immunohistochemistry studies formalin fixed paraffin embedded tissue sections are studied using the b sa system technique with appropriate positive a nd negative controls these ihc studies provide...		
0.018741	0.024572	0.889639	0.059386	0.000286	0.000286

Figure 8: BERT All-Fields Model Interpretation: Visualization of importance scores assigned to pathology report subfields outside of the diagnostic section for three separate pathology reports (A–C) that were assigned by raters CPT code 88360; information from report subfields that appear more red was utilized more by the model for the final prediction of the code; attention scores listed below the text from the subfields and title of each subfield supplied

interpretations of a few sample sentences for CPT codes using the *Diagnosis* BERT approach revealed important phrases that aligned with assignment of the respective CPT code [Figure 6C, D, and F]. Finally, we included a few examples of the attention mechanism used in the BERT approach, which highlights some of the many semantic/syntactic dependencies that the model finds within text subsections [Figure 7]. These attention matrices were plotted along with importance assigned to subsections of pathology reports using the *All-Fields* model [Figure 8], all with their respective textual content. Additional interpretation of reports for pathologists may be found in the Supplementary Materials (Supplementary Figures 13 and 14).

Discussion

In this study, we characterized a large corpus of almost 100,000 pathology reports at a mid-sized academic medical center. Our studies indicate that the XGBoost and BERT methodologies produce highly accurate predictions of both primary and ancillary CPT codes, which has the potential to save operating costs by first suggesting codes prior to manual inspection and flagging potential manual coding errors for review. Further, both the BERT and XGBoost models preserved the ordering of the code/case complexity, where most of the misclassifications were made between codes of a similar complexity. The model interpretations via SHAP suggest a terminology that is consistent with code complexity. For instance, "vulva," "uterus," and "adenocarcinoma" were associated with CPT code 88309. We noted associations between "endometrium diagnosis" and "esophagus" and CPT code 88305. "Biopsy" was associated with CPT codes 88305 and 88307, while "myocyte" was associated with CPT code 88307 (myocardium). In addition, we noticed a positive association between "products of conception" and lower complexity codes (CPT code 88304) and a negative association with higher complexity codes. The aforementioned associations uncovered using SHAP are consistent with reporting standards for histological examination.[31,32,57]

Previous studies predicting CPT codes have largely been unable to characterize the importance of different subsections of a pathology report. Using the BERT and XGBoost methods, we were also able to show that significant diagnostic / coding information is contained in nondiagnostic subsections of the pathology report, particularly the Clinical Information and Specimen Processing sections. Such information was more pertinent when predicting ancillary CPT codes, as nondiagnostic subfields are more likely to contain test ordering information, though performance gains were observed for primary codes when employing the XGBoost model over an entire pathology report. This is expected, as many of the CPT codes are based on procedure type / specimen complexity and ancillary CPT codes are expected to contain more informative text in the nondiagnostic sections. Potentially, the variable presence/absence of different reporting subfields may have made predicting primary codes using the BERT model more difficult, as the extraction of information different subsections was not optimized for aside from how much weight to apply to each section.

Although our prediction accuracy is comparable to previous reports of CPT prediction using machine-learning methods, our work covers a wider range of

codes than previously reported, compares the different algorithms through rigorous cross-validation, reports a significantly higher sensitivity and specificity, and demonstrates the importance of utilizing other parts of the pathology report for procedural code prediction. Further, previous works had only considered the first 100 words of the diagnostic section and had failed to properly account for class-balancing, potentially leading to inflated performance statistics; however, our study carefully considers the ordinality of the response and reports macro-averaged measures that take into account infrequently assigned codes.

We also demonstrated that the pathology report subfields contained pertinent diagnostic and procedural information that could adequately separate our text corpus based on ancillary CPT codes and the signing pathologist. With regard to ancillary testing, it was interesting to note how some of the clinical codes for acquisition and quantification of markers on specialized stains (CPT 88341, 88342, 88344, 88360) performed the worst overall, which may potentially suggest inconsistent reporting patterns for the ordering of specialized stains.[34] The revision of CPT codes 88342 and 88360, and the addition of CPT codes 88341 and 88344 in 2015 lay just outside of the range of the data collection period, which was from June 2015 to June 2020.[58] Evolving coding/billing guidelines will always present challenges when developing NLP guidelines for clinical tests, though our models' optimal performance and the fact that major coding changes occurred outside of the data collection period suggest that temporal changes in coding patterns did not likely impact the ability to predict CPT codes. We did not find significant changes in the assignment of most of the primary codes over the study period. Since major improvements were obtained through incorporating the other report subfields for the codes, nondiagnostic text may be more important for records of specialized stain processing and should be utilized as such.

Limitations

There are a few limitations to our work. For instance, due to computational constraints, most BERT models can only take as input 512 words at a time (Supplementary Section "Additional Information on BERT Pretraining"). We utilized a pretrained BERT model that inherited knowledge from large existing biomedical data repositories at the expense of flexibility in sequence length size (i.e. we could not modify the word limit while utilizing this pretrained model). We noticed that in our text corpus, less than 2% of reports were longer than this limitation and thus had to be truncated when input into the deep learning model, which may impact results. Potentially, longer pathology reports describe more complicated cases, which may utilize additional procedures. From our cluster analysis, we demonstrated that this appeared to be the case for a subset of report clusters, though for one cluster, the opposite was true. However, a vast majority of pathology reports fell within the BERT word limits, so we considered any word length-based association with CPT code complexity to have negligible impact on the model results. The XGBoost model, alternatively, is able to operate on the entire report text. Thus, XGBoost may more directly capture interactions between words spanning across document subsections pertaining to complex cases, which may serve as one plausible explanation of its apparent performance increase with respect to the BERT approaches. Although we attempted to take into account the ordinality of case complexity for the assignment of primary CPT codes, such work should be revisited as ordinal loss functions for both deep learning and tree-based models become more readily available. There were also cases where multiple primary codes were assigned; whereas the ancillary codes were predicted by using a multitarget objective, and the primary code prediction can be configured similarly though this was outside the scope of the study.[32] Although we conducted coarse hyperparameter scans, we note that generally such methods are deemed both practical and acceptable. Although other advanced hyperparameter scanning techniques exist (e.g., Bayesian optimization or genetic algorithm), in many cases, these methods obtain performance similar to randomized hyperparameter searches and may be far more resource intensive.[59]

Future directions

Given the secondary objectives of our study (e.g., prediction of ancillary codes, studying sources of variation in text, i.e. pathologist), we were able to identify additional areas for follow-up.

First, we were able to assess nuanced pathologist-specific language, which was largely determined by specialty (e.g. subspecialties such as cytology use highly regimented language, making it more difficult to separate practitioners). There is also potentially useful information to be gained by working to identify text that can distinguish pathologists within subspecialties (found as a flag in the Pathology Information System) and conditional on code assignment rather than identify pathologists across subspecialties. This information can be useful in helping to create more standardized lexicons / diagnostic rubrics (for instance, The Paris System for Urine Cytopathology[60]). Research into creating a standard lexicon for particular specialties or converting raw free text into a standardized report could be very fruitful, especially for the positive impact it would have in allowing nonpathologist physicians to more easily interpret pathology reports and make clinical decisions. As an example of how nonstandardized text lexicon can impact reporting, it has long been suspected that outlier text can serve as a marker of uncertainty or ambiguity about the diagnosis. For instance, if there is a text content

outlier in a body of reports with the same CPT code, then we can hypothesize that such text may be more prone to ambiguous phrases or hedging, from which pathologists may articulate their uncertainty for a definitive diagnosis. As such, we would also like to assess the impact of hedging in the assignment of procedural codes, and further its subsequent impact on patient care. As another example, excessive ordering of different specialized stains and pathology consults may suggest indecisiveness, as reflected in the pathology report. To ameliorate these differences in reporting patterns, generative deep learning methods can be employed to summarize the text through the generation of a standard lexicon.

Other excellent applications of BERT-based text models include the prediction of relative value units (RVU's) via report complexity for pathologist compensation calculations (which is related to primary code assignment) and the detection of cases that may have been mis-billed (e.g., a code of lower complexity was assigned), which can potentially save the hospital resources.[61] We are currently developing a web application that will both interface with the Pathology Information System and can be used to estimate the fiscal impact of underbilling by auditing reports with false positive findings. Tools such as *Inspirata* can also provide additional structuring for our pathology reports outside of existing schemas.[62]

Although much of the patient's narrative may be told separately through text, imaging, and omics modalities,[63] there is tremendous potential to integrate semantic information contained in pathologist notes with imaging and omics modalities to capture a more holistic perspective of the patient's health and integrate potentially useful information that could otherwise be overlooked. For instance, the semantic information contained in a report may highlight specific morphological and macro-architectural features in the correspondent biopsy specimen that an image-based deep learning model might struggle to identify without additional information. Although XGBoost demonstrated equivalent performance with the deep learning methods used for CPT prediction, its usefulness in a multimodal model is limited because these machine-learning approaches rely heavily on the feature extraction approach, where feature generation mechanisms using deep learning can be tweaked during optimization to complement the other modalities. Alternatively, the semantic information contained within the word embedding layers of the BERT model can be fine-tuned when used in conjunction with or directly predicting on imaging data, allowing for more seamless integration of multimodal information. Integrating such information, in addition to structured text extraction systems (i.e., named entity recognition) that can recognize and correct the mention of such information in the text, may provide a unique search functionality that can benefit experiment planning.[34]

Although comparisons between different machine-learning models may inform the optimal selection of tools that integrate with the Pathology Information System, we acknowledge that such comparisons can benefit from updating as new machine-learning architectures are developed. As such, we plan to incorporate newer deep learning architectures, such as the Reformer or Albert, which do not suffer from the word length limitations of BERT, though training all possible language models was outside of the scope of our study since pretrained medical word embeddings were not readily available at the time of modeling.

Conclusion

In this study, we compare three cutting-edge machine learning techniques for the prediction of CPT codes from pathology text. Our results provide additional evidence for the utility of machine-learning models to predict CPT codes in a large corpus of pathology reports acquired from a mid-sized academic medical center. Further, we demonstrated that utilizing text from parts of the document other than the diagnostic section aids in the extraction of procedural information. Although both the XGBoost and BERT methodologies yielded comparable results, either method can be used to improve the speed and accuracy of coding by the suggestion of relevant CPT codes to coders, though deep learning approaches present the most viable methodology for incorporating text data with other pathology modalities.

Acknowledgments

We would like to thank Matthew LeBoeuf for thoughtful discussion.

Financial support and sponsorship

This work was supported by NIH grants R01CA216265, R01CA253976, and P20GM104416 to BC, Dartmouth College Neukom Institute for Computational Science CompX awards to BC and LV, and Norris Cotton Cancer Center, DPLM Clinical Genomics and Advanced Technologies EDIT program. JL is supported through the Burroughs Wellcome Fund Big Data in the Life Sciences at Dartmouth. The funding bodies above did not have any role in the study design, data collection, analysis and interpretation, or writing of the manuscript.

Authors' contributions

The conception and design of the study were contributed by JL and LV. Initial analyses were conducted by JL and NV. All authors contributed to writing and editing of the manuscript and all authors read and approved the final manuscript.

Conflicts of interest

There are no conflicts of interest.

References

1. Mantas J, Hasman A. Informatics, Management and Technology in Healthcare. Amsterdam: IOS Press; 2013.
2. Wilson RA, Chapman WW, Defries SJ, Becich MJ, Chapman BE. Automated ancillary cancer history classification for mesothelioma patients from free-text clinical reports. J Pathol Inform 2010;1:24.
3. Sheikhalishahi S, Miotto R, Dudley JT, Lavelli A, Rinaldi F, Osmani V. Natural language processing of clinical notes on chronic diseases: Systematic review. JMIR Med Inform 2019;7:e12239.
4. Assale M, Dui LG, Cina A, Seveso A, Cabitza F. The revival of the notes field: Leveraging the unstructured content in electronic health records. Front Med (Lausanne) 2019;6:66.
5. Spasic I, Nenadic G. Clinical text data in machine learning: Systematic review. JMIR Med Inform 2020;8.
6. Horsky J, Drucker EA, Ramelson HZ. Accuracy and completeness of clinical coding using ICD-10 for ambulatory visits. AMIA Annu Symp Proc 2017;2017:912-20.
7. Senders JT, Cote DJ, Mehrtash A, Wiemann R, Gormley WB, Smith TR. Deep learning for natural language processing of free-text pathology reports: A comparison of learning curves. BMJ Innovations 2020;6:192-8.
8. Savova GK, Danciu I, Alamudun F, Miller T, Lin C, Bitterman DS, *et al*. Use of natural language processing to extract clinical cancer phenotypes from electronic medical records. Cancer Res 2019;79:5463-70.
9. Alawad M, Hasan SMS, Christian JB, Tourassi G. Retrofitting word embeddings with the UMLS metathesaurus for clinical information extraction. In 2018 IEEE International Conference on Big Data (Big Data), 2018; 2838-46.
10. Levis M, Leonard Westgate C, Gui J, Watts BV, Shiner B. Natural language processing of clinical mental health notes may add predictive value to existing suicide risk models. Psychol Med 2020;51:1-10.
11. Ghassemi M, Naumann T, Schulam P, Beam AL, Chen IY, Ranganath R. A review of challenges and opportunities in machine learning for health. AMIA Jt Summits Transl Sci Proc 2020;2020:191-200.
12. Weng WH, Wagholikar KB, McCray AT, Szolovits P, Chueh HC. Medical subdomain classification of clinical notes using a machine learning-based natural language processing approach. BMC Med Inform Decis Mak 2017;17:155.
13. Blei DM, Ng AY, Jordan MI. Latent Dirichl*et Al*location. J Mach Learn Res 2003;3:993-1022.
14. Ramos J. Using TF-IDF to Determine Word Relevance in Document Queries. Proceedings of the first Instructional Conference on Machine Learning, Association for Computing Machinery, New York, NY. 2003;242;133-42.
15. Nagendran M, Chen Y, Lovejoy CA, Gordon AC, Komorowski M, Harvey H, *et al*. Artificial intelligence versus clinicians: Systematic review of design, reporting standards, and claims of deep learning studies. BMJ 2020;368:m689.
16. Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M. Scalable and accurate deep learning with electronic health records. NPJ Digital Medicine 2018;1:1-10.
17. Shickel B, Tighe PJ, Bihorac A, Rashidi P. Deep EHR: A survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. IEEE J Biomed Health Inform 2018;22:1589-604.
18. Devlin J, Chang M-W, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, MN: Association for Computational Linguistics, 2019. p. 4171-86.
19. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, *et al*. Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. NIPS'17. Red Hook, NY: Curran Associates.; 2017. p. 6000-10.
20. Qiu J, Yoon H-J, Oak RNL (ORNL), Fearn PA, Tourassi GD. Deep learning for automated extraction of primary sites from cancer pathology reports. IEEE J Biomed Health Inform 2017;22;244-51.
21. Gao S, Young MT, Qiu JX, Yoon HJ, Christian JB, Fearn PA, *et al*. Hierarchical attention networks for information extraction from cancer pathology reports. J Am Med Inform Assoc 2018;25:321-30.
22. Buckley JM, Coopey SB, Sharko J, Polubriaginof F, Drohan B, Belli AK, *et al*. The feasibility of using natural language processing to extract clinical information from breast pathology reports. J Pathol Inform 2012;3:23.
23. Oliwa T, Maron SB, Chase LM, Lomnicki S, Catenacci DVT, Furner B, *et al*. Obtaining knowledge in pathology reports through a natural language processing approach with classification, named-entity recognition, and relation-extraction heuristics. JCO Clin Cancer Inform 2019;3:1-8.
24. Arnold CW, El-Saden SM, Bui AA, Taira R. Clinical case-based retrieval using latent topic analysis. AMIA Annu Symp Proc 2010;2010:26-30.
25. Kalra S, Li L, Tizhoosh HR. Automatic Classification of Pathology Reports using TF-IDF Features. arXiv:190307406 [cs, stat] March 2019.
26. Xu K, Lam M, Pang J, Gao X, Band C, Mathur P, *et al*. Multimodal Machine Learning for Automated ICD Coding. In: Machine Learning for Healthcare Conference. PMLR; 2019. p. 197-215.
27. Saib W, Chiwewe T, Singh E. Hierarchical deep learning classification of unstructured pathology reports to automate ICD-O morphology grading. arXiv:200900542 [cs] August 2020.
28. Ye JJ. Construction and utilization of a neural network model to predict current procedural terminology codes from pathology report texts. J Pathol Inform 2019;10:13.
29. Dotson P. CPT® codes: What are they, why are they necessary, and how are they developed? Adv Wound Care (New Rochelle) 2013;2:583-7.
30. Hanauer DA, Liu Y, Mei Q, Manion FJ, Balis UJ, Zheng K. Hedging their mets: The use of uncertainty terms in clinical documents and its potential implications when sharing the documents with patients. AMIA Annu Symp Proc 2012;2012:321-30.
31. Deeken-Draisey A, Ritchie A, Yang GY, Quinn M, Ernst LM, Guttormsen A, *et al*. Current procedural terminology coding for surgical pathology: A review and one academic center's experience with pathologist-verified coding. Arch Pathol Lab Med 2018;142:1524-32.
32. Dimenstein IB. Principles and controversies in CPT coding in surgical pathology. Lab Med 2011;42:242-9.
33. Joo H, Burns M, Kalidaikurichi Lakshmanan SS, Hu Y, Vydiswaran VGV. Neural machine translation-based automated current procedural terminology classification system using procedure text: Development and validation study. JMIR Form Res 2021;5:e22461.
34. Ye JJ. Using an R program to monitor pathology reports for omissions in reporting ancillary tests and errors in test names. Arch Pathol Lab Med 2020;144:917-8.
35. Milinovich A, Kattan MW. Extracting and utilizing electronic health data from epic for research. Ann Transl Med 2018;6:42.
36. Bosker HR. Using fuzzy string matching for automated assessment of listener transcripts in speech intelligibility studies. Behav Res Methods 2021;53:1945-53.
37. Neamatullah I, Douglass MM, Lehman LW, Reisner A, Villarroel M, Long WJ, *et al*. Automated de-identification of free-text medical records. BMC Med Inform Decis Mak 2008;8:32.
38. Montani I, Honnibal M, Honnibal M, Landeghem SV, Boyd A, Peters H, *et al*. SpaCy: Industrial-Strength Natural Language Processing in Python. Zenodo; 2021.
39. McInnes L, Healy J, Saul N, Großberger L. UMAP: Uniform manifold approximation and projection. J Open Source Softw 2018;3:861.

40. McInnes L, Healy J, Astels S. HDBSCAN: Hierarchical density based clustering. J Open Source Softw 2017;2:205.
41. Bonett DG. Point-biserial correlation: Interval estimation, hypothesis testing, meta-analysis, and sample size determination. Br J Math Stat Psychol 2020;73:113-44.
42. Platt JC. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In: Smola AJ, Bartlett P, Schölkopf B, Schuurmans D, editors. Advances in Large Margin Classifiers. Cambridge, MA: MIT Press; 1999. p. 61-74.
43. Hearst M, Dumais ST, Osman E, Platt J, Scholkopf B. Support vector machines. IEEE Intel Syst Appl 1998;13:18-28.
44. Wen Z, Shi J, Li Q, He B, Chen J. ThunderSVM: A fast SVM library on GPUs and CPUs. J Mach Learn Res 2018;19:1-5.
45. Chen T, Guestrin C. XGBoost: A scalable tree boosting system. In: Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '16. New York, NY: ACM; 2016. p. 785-94.
46. Loh W-Y. Classification and regression trees. Wiley Interdiscip Rev Data Min Knowl Discov 2011;1:14-23.
47. Breiman L. Random forests. Mach Learn 2001;45:5-32.
48. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:436-44.
49. Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, *et al*. Transformers: State-of-the-Art natural language processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. Online: Association for Computational Linguistics, 2020. p. 38-45.
50. Alsentzer E, Murphy J, Boag W, Weng W-H, Jindi D, Naumann T, *et al*. Publicly available clinical BERT embeddings. In: Proceedings of the 2nd Clinical Natural Language Processing Workshop. Minneapolis, MN: Association for Computational Linguistics; 2019. p. 72-8.
51. McCullagh P. Proportional odds model: Theoretical background. In: Wiley StatsRef: Statistics Reference Online. Hoboken, NJ: American Cancer Society; 2014.
52. Khattak FK, Jeblee S, Pou-Prom C, Abdalla M, Meaney C, Rudzicz F. A survey of word embeddings for clinical text. J Biomed Inform 2019;100S:100057.
53. Johnson AE, Pollard TJ, Shen L, Lehman LW, Feng M, Ghassemi M, *et al*. MIMIC-III, a freely accessible critical care database. Sci Data 2016;3:160035.
54. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, *et al*. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell 2020;2:56-67.
55. Kokhlikyan N, Miglani V, Martin M, Wang E, Alsallakh B, Reynolds J, *et al*. Captum: A unified and generic model interpretability library for PyTorch. arXiv:200907896 [cs, stat] September 2020.
56. Sundararajan M, Taly A, Yan Q. Axiomatic attribution for deep networks. In International Conference on Machine Learning. PMLR. Toulon, France. 2017;5;3319-28.
57. Bonert M, Zafar U, Maung R, El-Shinnawy I, Kak I, Cutz JC, *et al*. Evolution of anatomic pathology workload from 2011 to 2019 assessed in a regional hospital laboratory via 574,093 pathology reports. PLoS One 2021;16:e0253876.
58. A Look Ahead: Pathology CPT Changes for 2015 | APS Medical Billing. Available from: https://apsmedbill.com/whitepapers/look-ahead-pathology-cpt-changes-2015. [Last accessed on 2021 Feb 11].
59. Mayhew MB, Tran E, Choi K, Midic U, Luethy R, Damaraju N, *et al*. Optimization of genomic classifiers for clinical deployment: Evaluation of Bayesian optimization to select predictive models of acute infection and in-hospital mortality. Pac Symp Biocomput 2021;26:208-19.
60. Vaickus LJ, Suriawinata AA, Wei JW, Liu X. Automating the Paris system for urine cytopathology-A hybrid deep-learning and morphometric approach. Cancer Cytopathol 2019;127:98-115.
61. Kim Y, Lee JH, Choi S, Lee JM, Kim JH, Seok J, *et al*. Validation of deep learning natural language processing algorithm for keyword extraction from pathology reports in electronic health records. Sci Rep 2020;10:20265.
62. Cernile G, Heritage T, Sebire NJ, Gordon B, Schwering T, Kazemlou S, *et al*. Network graph representation of COVID-19 scientific publications to aid knowledge discovery. BMJ Health Care Inform 2021;28:e100254.
63. Frolov S, Hinz T, Raue F, Hees J, Dengel A. Adversarial text-to-image synthesis: A review. Neural Netw 2021;144:187-209.

Supplementary Materials

Supplementary Methods

Additional Information on Deidentification Approach

In this section, we have compiled a list of all publicly available datasets used to remove identifiable patient names from the report text:

1. First Names
    a. https://github.com/ankane/age/blob/master/names/
    b. https://github.com/smashew/NameDatabases/blob/master/NamesDatabases/first%20names/all.txt
    c. https://hackage.haskell.org/package/gender
    d. https://raw.githubusercontent.com/solvenium/names-dataset/master/dataset/Male_given_names.txt
    e. https://raw.githubusercontent.com/solvenium/names-dataset/master/dataset/Female_given_names.txt
    f. https://github.com/philipperemy/name-dataset/tree/master/names_dataset/v1
2. Last Names
    a. https://github.com/smashew/NameDatabases/blob/master/NamesDatabases/surnames/all.txt
    b. https://raw.githubusercontent.com/solvenium/names-dataset/master/dataset/Surnames.txt
    c. https://github.com/philipperemy/name-dataset/tree/master/names_dataset/v1

Additional Description of Topic Modeling and Report Characterization Techniques: TF-IDF, UMAP, HSBSCAN, LDA

Here, we briefly provide an overview of the modeling techniques that, when utilized in conjunction, characterized the pathology report corpus through the establishment of important words that were not ubiquitous across the corpus (TF-IDF), removal of noise and discovery of clusters (UMAP and HDBSCAN), and generating topics that describe recurrent themes (LDA).

TF-IDF (term frequency inverse document frequency) takes as input a sparse count matrix, which contains the reports as rows and individual words/n-grams as columns, where each element is a count of the n-gram in the document. TF-IDF re-weights the count matrix based on an algorithm that modifies word importance on the basis of whether the word is ubiquitous across all documents and/or enriched in its own document. The formula for TF-IDF is:

$$tf - idf_{t,d} = tf_{t,d} * \log \frac{n}{1 + df_t} + 1$$

where t and d refer to the specific term and document, respectively. The term-frequency, tf, is the reported count of the n-gram in the particular document, whereas document frequency, df, is the number of reports that contain the term (i.e. how ubiquitous the word is across the corpus). Usually these values are normalized via the euclidean norm to downweight longer documents.

Such information may replace the count matrix for downstream analysis, though it is not necessary.

The UMAP operates on the count/tf-idf matrix to reduce the dimensionality of the reports while preserving the key relationships between the reports. This is unlike PCA, which selects principal components to maximize variance, and TSNE (T-Stochastic Neighborhood Embedding), which learns a lower dimensional manifold that preserves local distance between reports. The UMAP forms fuzzy simple sets that represent the higher dimensional manifold at multiple distances. Computationally, this amounts to constructing a weighted nearest-neighbors graph and an optimization routine that preserves a similar structure in the low-dimensional manifold while optimizing a force-directed graph layout.

HDBSCAN is a clustering algorithm that operates on the lower dimensional manifold to find natural groupings of the data. HDBSCAN combines hierarchical clustering techniques, which iteratively merge similar clusters, with density-based clustering, which estimates clusters of a similar density. HDBSCAN estimates the density of points based on whether a certain number of points exist within a small well-defined neighborhood and whether two points share a common neighbor, both outside of that which is expected if there were noise. HDBSCAN varies the size of this neighborhood to consider/integrate density on multiple scales to form a hierarchy, which may be further processed to yield the clusters. This yields a set of clusters and points that have been defined as noise. Since the algorithm considers the notion of distance and connectedness on multiple scales / neighborhoods, it often pairs well with UMAP due to similarities in formulation.

Latent Dirichlet Allocation (LDA) is a three-level probabilistic/Bayesian generative model that is used for inferring a distribution of topics across a document corpus. Ultimately, the goal of the model is to provide a mechanistic model for how the count matrix arises (that is, estimating the frequency of words in each of the reports). The simplified conceptual framework is as follows:

1. A document is selected.
2. N number of words are selected from a Poisson distribution (which iterates steps 3-4 N times).
3. For each word (not yet selected), a topic is selected from a set of latent topics (topic mixture) that characterize the document with some probability.
4. A word is selected from a set of words that are ascribed to the topic.

5. The distribution of words selected via the generative approach in steps 1-3 are compared with the true distribution of words after marginalizing over the topics and documents.

The generative model initially places two separate Dirichlet priors over selecting topics (topic mixture) and words from topics (a k words by V topics matrix). Variational Bayes and expectation maximization techniques are applied to estimate the posterior distribution of the topic mixture and topic-word parameters by assuming the conditional posterior follows a known family of distributions. Ultimately, sampling the predictive posterior allows for an inference of the distribution of topics across documents.

Additional Information on Hyperparameter Scans

We performed coarse hyperparameter searches for ideal model specifications. We registered optimal hyperparameters based on the loss over each validation set (alternatively by either an F1-score or an AUC metric), depending on the modeling approach. Model convergence was monitored by using the validation set; the test set was completely held out from the updating of parameters or the tuning of hyperparameters. Here, we list the hyperparameters scanned over for each model through coarse inspection of validation set statistics. Selection of this grid was based on a mixture of sensible recommendations and experimentation. Selected hyperparameters are marked in bold, and unlisted hyperparameters were set to package defaults:

- Support Vector Machine:
  - Kernel: **Radial Basis (RBF)**, Linear
  - Gamma (scales RBF distance): **Automatic** (set to [number features]$^{-1}$, where number features is 6 based on UMAP embeddings), 1, or 5
- XGBoost:
  - Max Depth: 2, 5, **8**, None (runs until split criterion are satisfied, e.g., minimum samples to split on)
  - Number of trees: 100, **300**, 600, 800
  - Number of GPU histogram bins (for optimal run time using GPU): 500, **800**, 2000
- BERT-Dx (Fine-tuning pretrained BERT model) (AdamW optimizer):
  - Batch size: **16**, 64
  - Number of Epochs: 1, **2**, 3, 5, which is typical for fine-tuning a BERT model
- BERT-All Fields (Adam optimizer):
  - Learning rate: 1e-2, 1e-3, **5e-4**, 1e-4, 1e-5
  - Batch size: 4, **8**, 16, 64
  - Number of epochs: 25, **100**
  - Loss function: **Cross-Entropy**, Ordinal Penalized Cross-Entropy

We note here that the BERT All-Fields model was trained by using a cosine annealing learning rate scheduler, which oscillates between the selected chosen rate and a η_{min} value of 1e-5 repetitively over the course of many epochs. This serves to scan a range of potential learning rates for optimal validation loss, from which to terminate training. Similarly, the BERT-Dx model was trained with an initial learning rate of 5e-5 for fine-tuning, with a linear decay scheduler, from which the learning rate asymptotically decreased toward zero. The BERT-Dx model was fine-tuned to predict specific code(s)/pathologist(s) and update pretrained word embeddings for input to the BERT-All Fields model. We tested an ordinal loss function that penalized misclassifications by adjacent categories/code complexity less than more distant codes. Weight decay was employed for both the BERT-Dx and All-Fields models as additional regularization.

Additional Information on BERT Pretraining

The BERT-Dx and BERT All-Fields models were pretrained by using the Bio-ClinicalBERT model, of which details for pre-training can be found here: https://huggingface.co/emilyalsentzer/Bio_ClinicalBERT. These word embeddings were downloaded a total of 1,290,981 times in the month of November 2021 alone, which demonstrates the widespread adoption of such an embedding system despite the fixed length nature of BERT inputs (512 words per input sequence). This embedding system was adopted for the purpose of this study because at the time of adoption, these pretraining embeddings were the most widely adopted embedding system and were only available for BERT. Further, leveraging a publicly available set of embeddings provides added value and generalization for fine-tuning our in-house dataset versus training from scratch. As additional evidence, other nondiagnostic report subfields were not fine-tuned on our corpus and performed well using the *All-Fields* models using the public embeddings alone, and only a small fraction of our corpus featured reports with a length that exceeded 512 words.

Additional Description of Explanation Techniques: SHAP, Integrated Gradients, Self-Attention, Attention Over Pathology Report Subfields

Shapley additive explanations (SHAP) is a technique that explains the results of any machine-learning model, which may have a complex decision surface. SHAP approximates this surface on a sample-by-sample basis by fitting one local additive model per sample. The coefficients of this model represent the importance of a feature or word. Local additive models operate to directly estimate the prediction of the machine-learning model when summed. That is, if f is the machine-learning model, g, for pathology

report i, with term frequency x, then the approximation is as follows:

$$f(x_i) \approx g_i(x_i) = E[f(z)] + \sum_k \phi_{k,i}(f, x_i)$$

Here, $\phi_{k,i}$ represents the shapley coefficient for term k of report i. The fitting procedure decides how to distribute the remainder between the mean value of the learned model over the dataset and the prediction to each of the predictors, while considering the importance of the individual predictor over the permutation or ensemble of the possible orderings of predictors when assigning reward (remainder). The predictor importance derived for individual CPT codes or pathologists was estimated by averaging these term/word-level importances / shapley coefficients across the entire document (we subsampled with a random seed for a more efficient computation) report corpus for a given model. This analysis was conducted for the XGBoost modeling approaches.

We utilized integrated gradients to interpret which words were found to be important for individual sentences when utilizing the BERT model, which we applied on the diagnostic text. Integrated gradient is a backpropagation-based method that is used for identifying salient features. Many traditional methods for ascertaining important predictors will take the gradient of the model prediction with a defined input $\nabla F(\vec{x})$, which serves as a linear approximation to the complex functional approximation, and multiply by the original input, $\vec{x}$, to yield the predictor specific importance $(\vec{x} \odot \nabla F(\vec{x}))$. However, this is less than ideal when $\vec{x}$ exists in the domain where the gradient saturates and also has no baseline for comparison. Integrated gradients, related to shapley values, overcome these two issues by first establishing a noninformative baseline/counterfactual $\vec{x_0}$; then, they successively sum more informative gradients along the path from the baseline to the observation $\vec{x_i}$ to yield the overall importance of the predictors:

$$IG(\vec{x_i}) = (\vec{x_i} - \vec{x_0}) * \int_{\alpha=0}^{\alpha=1} \nabla F(\vec{x_0} + \alpha(\vec{x_i} - \vec{x_0}))$$

Much of the success of the BERT methodology can be attributed to a neural network modeling approach known as the Transformer. As input to the model, each word is mapped to a semantic vector that captures the word's meaning, which is updated throughout the training process. The Transformer contextualizes the set of word vectors in a report through its encoder and decoder layers. The encoder and decoder layers are further decomposed into self-attention and feed-forward neural networks. Self-attention mechanisms capture dependencies between words within the sentence by forming a weight between each word and individually all of the words of the sentence; that is, identifying the most relevant words for the understanding of the current word. This is accomplished by estimating a weight between two words of a sentence. Here, matrix operations may be employed to speed up the calculation of the self-attention.

Suppose the word embeddings of the sentence are encapsulated in matrix $\overleftrightarrow{X}$, where rows indicate words and columns are the latent dimensions. Parameterized query, key, and value matrices are generated via the following operations:

$$\overleftrightarrow{Q} = W_Q \overleftrightarrow{X}$$

$$\overleftrightarrow{K} = W_K \overleftrightarrow{X}$$

$$\overleftrightarrow{V} = W_V \overleftrightarrow{X}$$

The query and key vectors are utilized as follows to construct the paired attention weights across a sentence, which could be thought of as learning/estimating a weighted unipartite matrix, attention matrix $\overleftrightarrow{A}$ (d_k is used for further normalization):

$$\overleftrightarrow{A} = softmax\left(\frac{\overleftrightarrow{Q}\overleftrightarrow{K}^T}{\sqrt{d_k}}\right)$$

$\overleftrightarrow{A}$ is the estimate of the word-to-word dependencies in the sentence for this particular operation. The embeddings of the sentence are updated/contextualized by multiplying this self-attention matrix with the embedding values:

$$\overleftrightarrow{Z} = \overleftrightarrow{A}\overleftrightarrow{V}$$

Usually, these self-attention matrices represent particular dependencies within the sentence. However, there may exist many complex dependencies to build a global understanding of the sentence/paragraph/report. As such, multiple self-attention "heads" are generated by allowing the existence of many query, key, and value matrices per encoding layer. We visualized the output of the estimated self-attention matrices in our article to demonstrate some of the learned dependencies. We have also omitted from this discussion nuanced specifics pertaining to the decoder (eg. retaining the query and key matrices from the encoder layers), residual connections, and positional embeddings, as they do not necessarily pertain to methods to interpret the output of the BERT model for a pathology report.

Attention across document subfields is entirely separate from BERT self-attention mechanisms. As mentioned in the main text, attention weights are utilized to decide how much information from report subsections to incorporate into the final global representation of the report. A weight matrix W, an n_z (number of latent dimensions) by one matrix (alternatively substituted by a gating neural network f_{gate}, as detailed in the main text), serves as a filter/gate to score how important a subsection is. The scores for each of the report subfields are softmaxed to assign a probability to each subfield for incorporation, $\vec{\alpha}$. The gate is learned via model parameter updates during

backpropagation. Importantly, we report the attention weights $\vec{\alpha}$ to communicate the importance of specific report subsections.

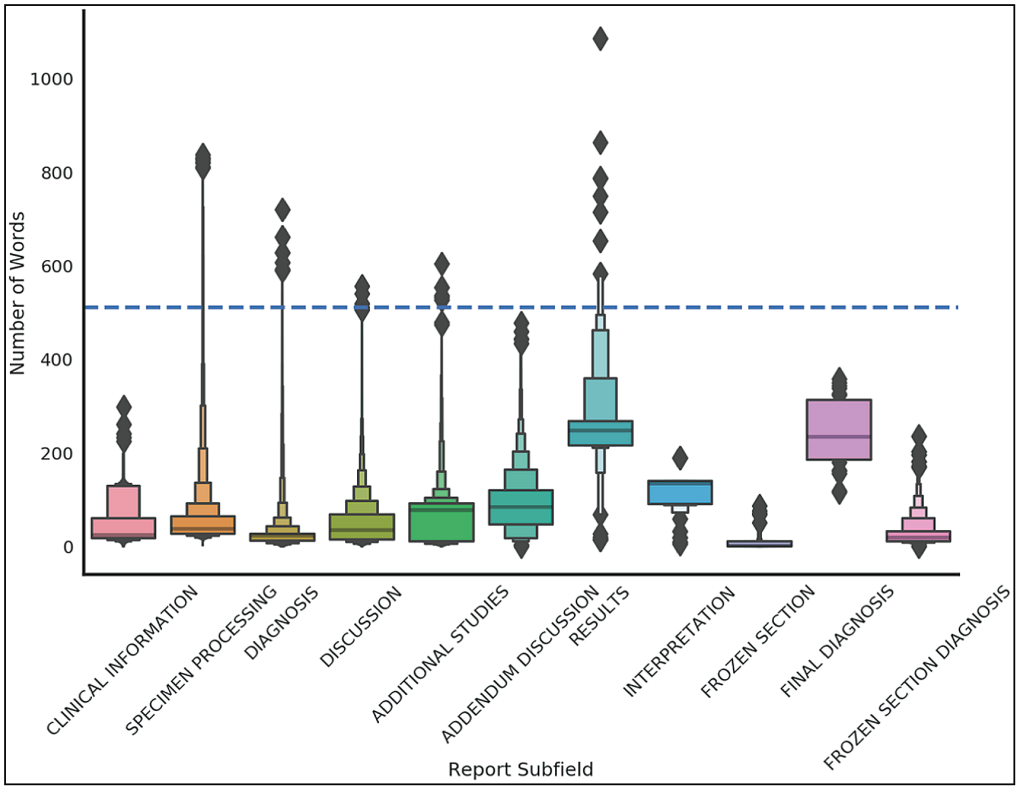
Supplementary Results

Supplementary Ancillary CPT Code Prediction Results

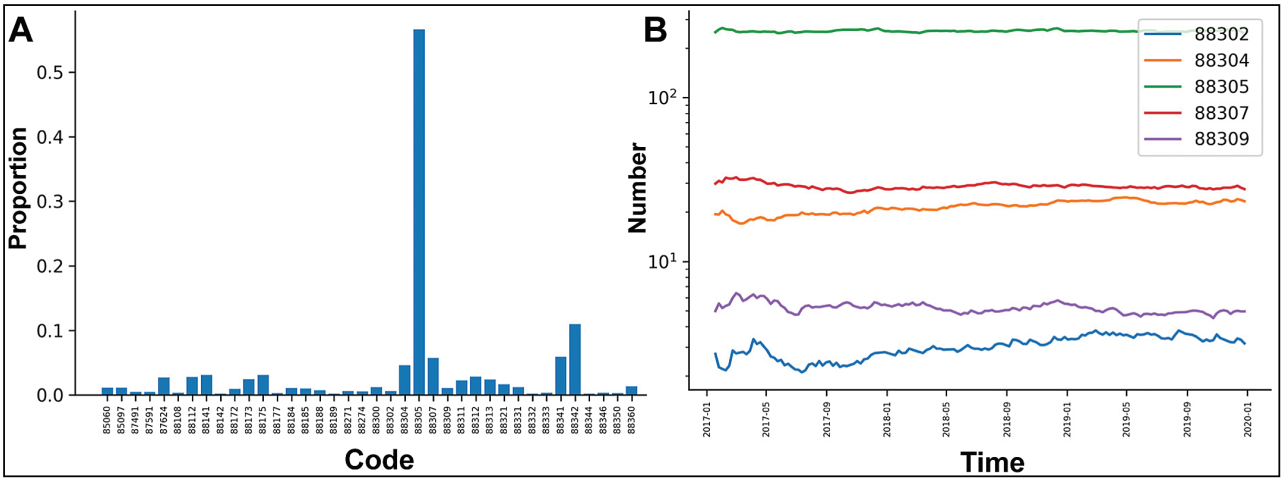
XGBoost (median AUC=0.997) outperformed BERT (median AUC=0.995) statistically ($p<0.001$) when utilizing all of the report subfields but given the high predictive performance these differences were not meaningful. Plots and tabulated statistics of the Youden Index derived from sensitivity/specificity of these algorithms across all of the validation folds confirm that utilizing information from all report subfields is better than utilizing information from the diagnostic text for the ancillary codes (Supplementary Table 10; Supplementary Figures 8–10). Averaging Youden's J statistic across all XGBoost and deep learning models, codes for immunohistochemistry/cytochemistry (CPT 88341, 88342, 88344, 88360), surgical pathology (CPT 88305), and flow cytometry (CPT 88188, 88189) performed worse versus other ancillary procedural codes; however, the performance improved considerably when including all report subfields for these codes (Supplementary Tables 6–9). Interestingly, the code for cytogenetic testing (CPT 88271) also experienced large improvements in sensitivity and specificity by incorporating other report subfields (Supplementary Table 10).

Supplementary Pathologist Prediction Results

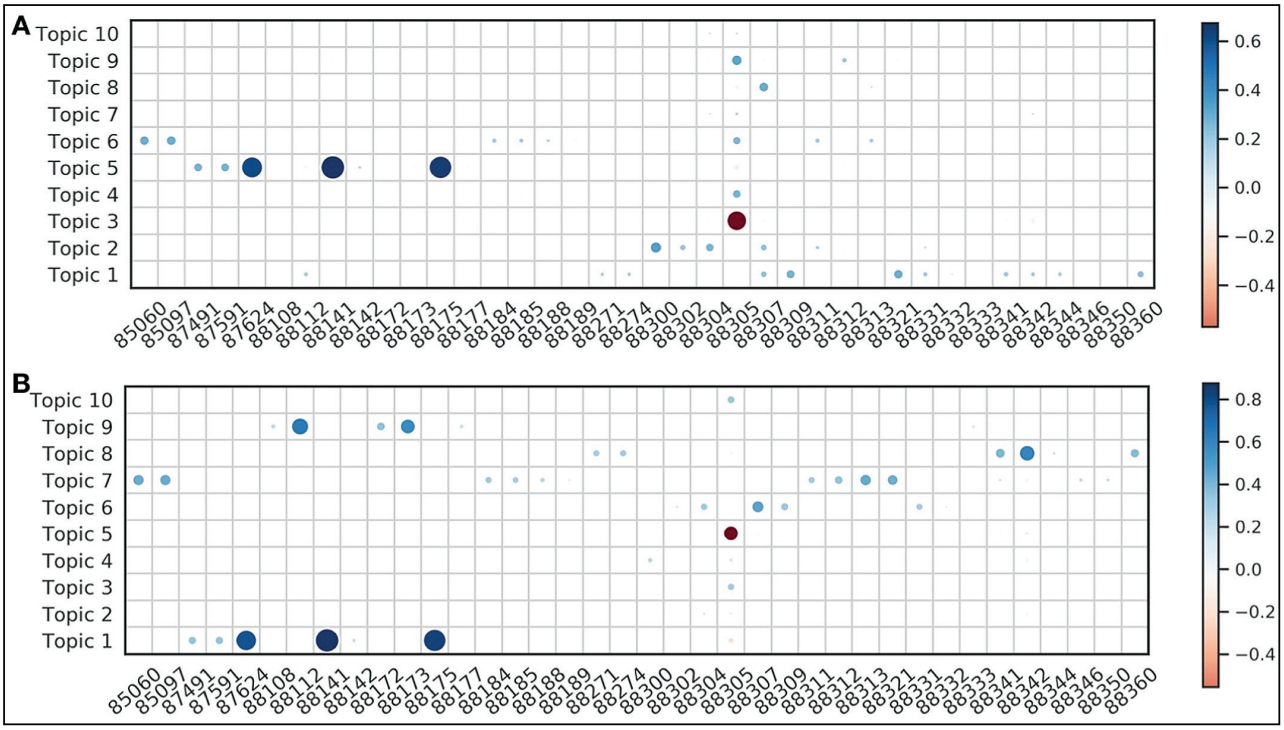
After subsetting to 64,583 documents that correspond to the 20 pathologists with the most sign-outs, the prediction of the pathologist who had written each pathology report was done with a reasonably high accuracy for the XGBoost and BERT approaches. BERT (macro-f1=0.72) performed comparably to XGBoost (macro-f1=0.71) for the prediction of pathologists on the diagnostic text; BERT (macro-f1=0.77) and XGBoost (macro-f1=0.78) also performed comparably when considering all report subfields (*all-fields*) (Supplementary Figure 11). Model performance improved when incorporating all report subsections. Interestingly, these pathologist-specific subtleties could not be distinguished via the SVM approach (Supplementary Tables 4 and 8). Comparisons between the embeddings formed by the *All-Fields* model and those using UMAP (Supplementary Figure 13A-B) show how the BERT methodology is able to extract features that are more pathologist specific, as compared with utilizing a bag-of-words approach. Comparing which pathologists were misclassified via the confusion matrix (Supplementary Figure 7 B) and corroboration with cross-tabulations with procedural codes (Supplementary Figure 7 A) demonstrates that pathologists with similar subspecialties were less distinguishable; however, individual patterns persist. We visualized some of the patterns that BERT was able to find in sample sentences via Integrated Gradients and important words via the XGBoost for select pathologists using SHAP (Supplementary Figure 14).



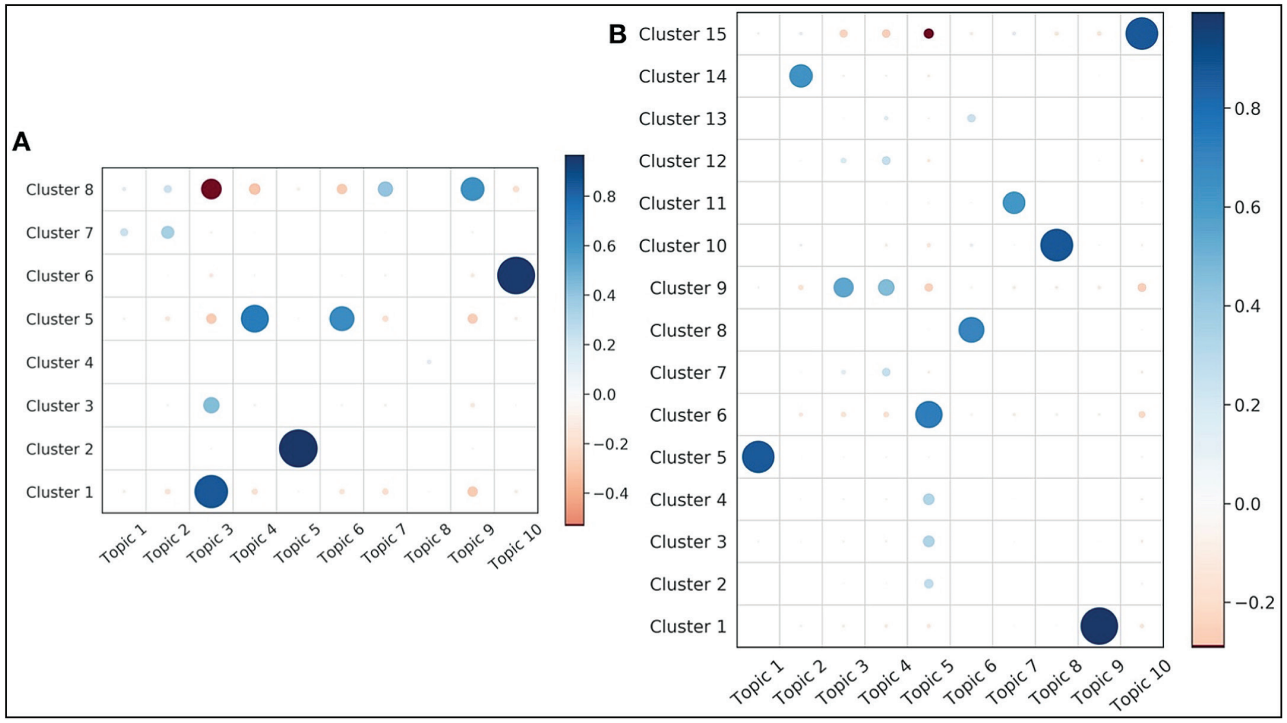
Supplementary Figure 1: Boxenplots of the number of words for each subfield across pathology report corpus; BERT cutoff word count of 512 words represented by a horizontal dashed line



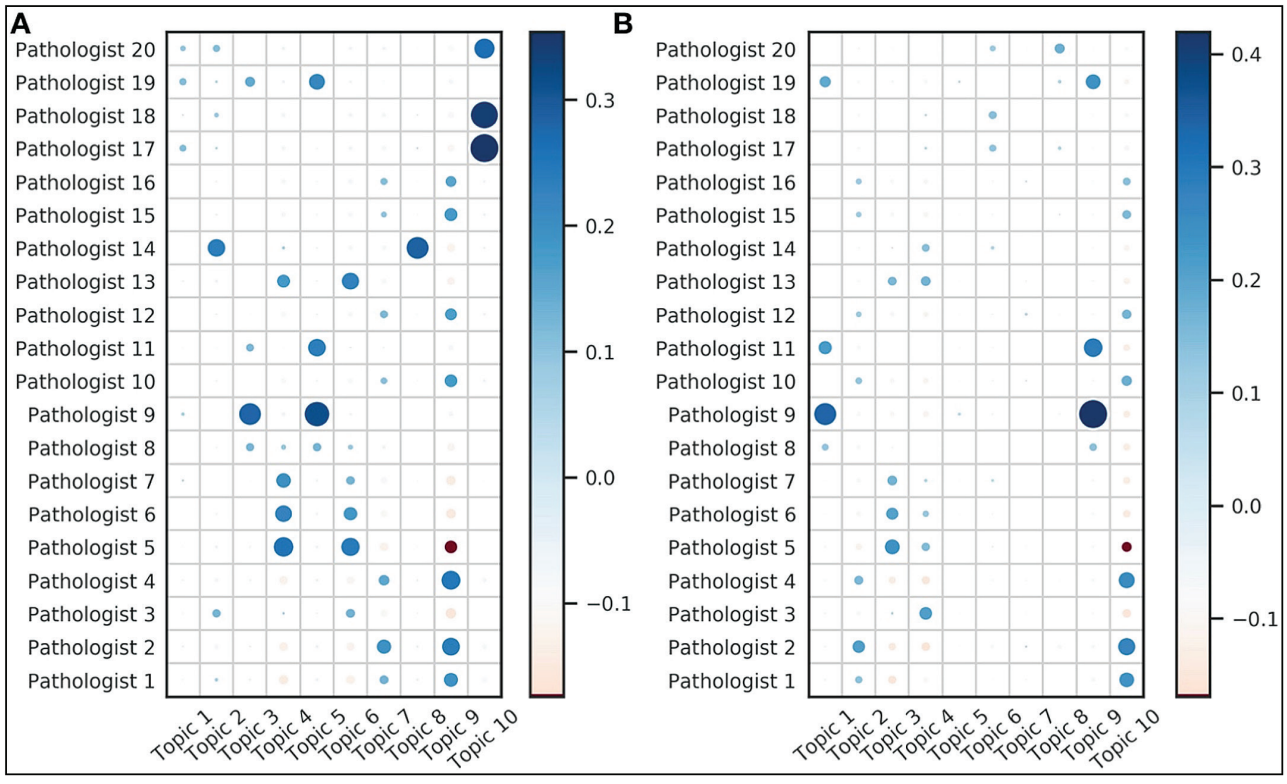
Supplementary Figure 2: CPT Code Statistics: A) Bar chart representing breakdown of the corpus by assigned codes (proportion); **B)** Changes in primary CPT codes over time, from 2017-2020, aggregated counts by week



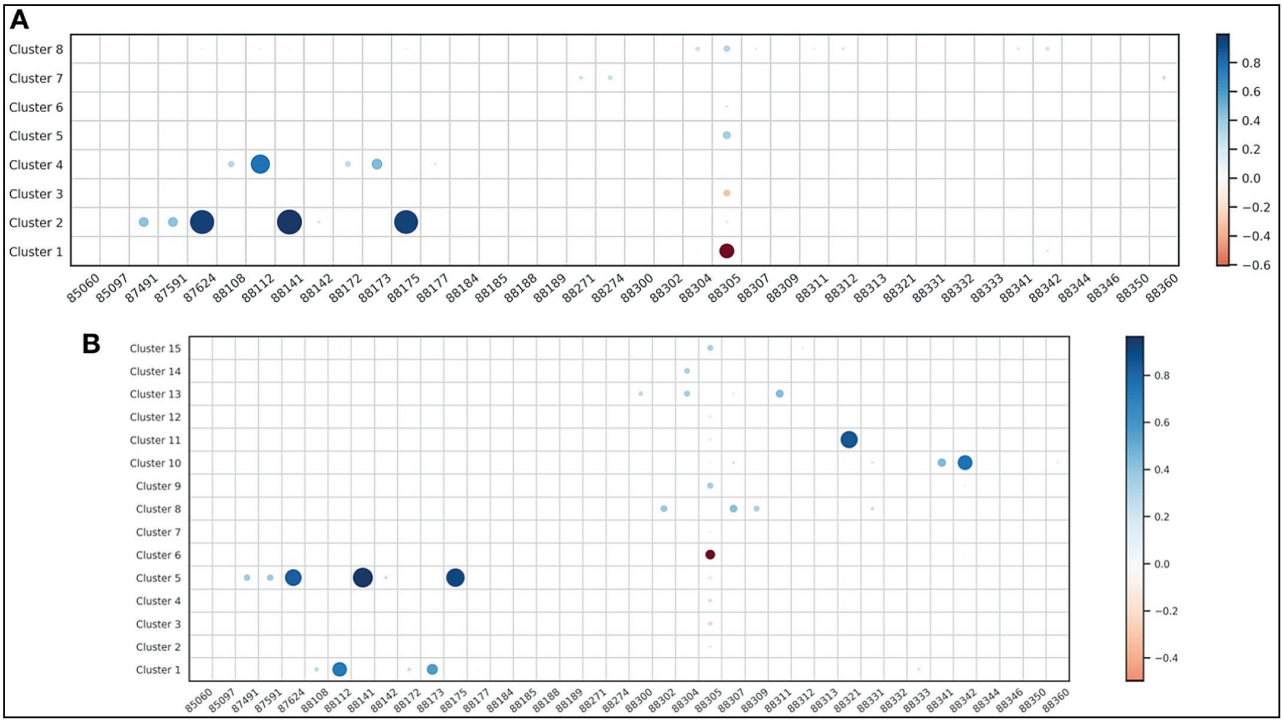
Supplementary Figure 3: Strength of correlation between topics and CPT codes denoted by the size and color of each circle; large blue circles indicate strong positive associations, whereas large red circles indicate strong negative associations; associations for: **A)** diagnostic text; **B)** all-fields text



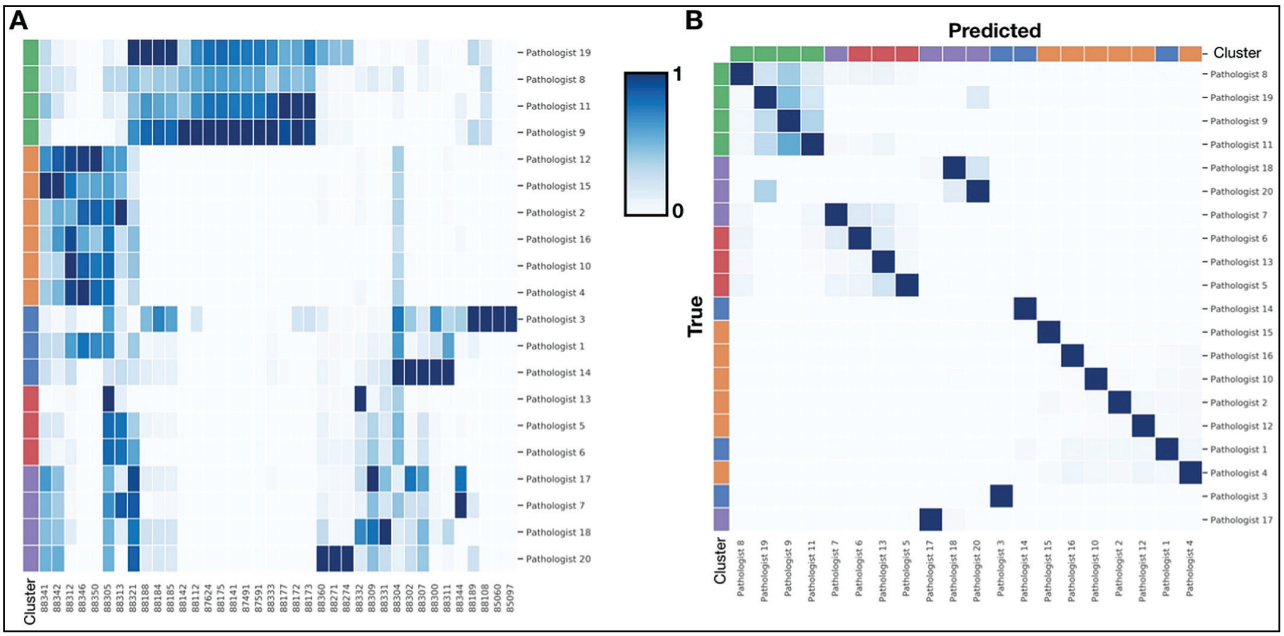
Supplementary Figure 4: Strength of correlation between topics and HDBSCAN report clusters denoted by the size and color of each circle; large blue circles indicate strong positive associations, whereas large red circles indicate strong negative associations; associations for: **A)** diagnostic text; **B)** all-fields text



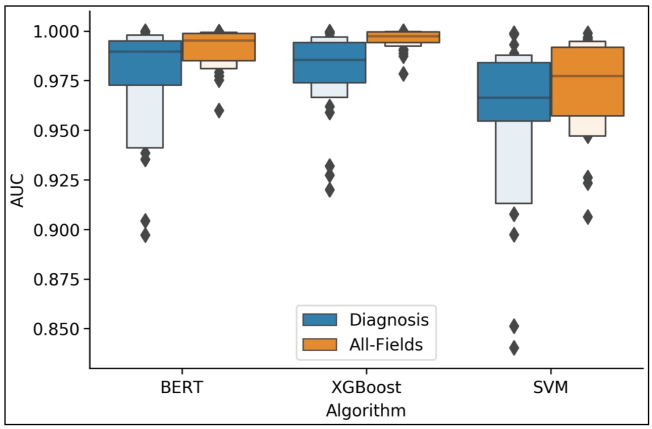
Supplementary Figure 5: Strength of correlation between topics and individual pathologists denoted by the size and color of each circle; large blue circles indicate strong positive associations, whereas large red circles indicate strong negative associations; associations for: **A)** diagnostic text; **B)** all-fields text



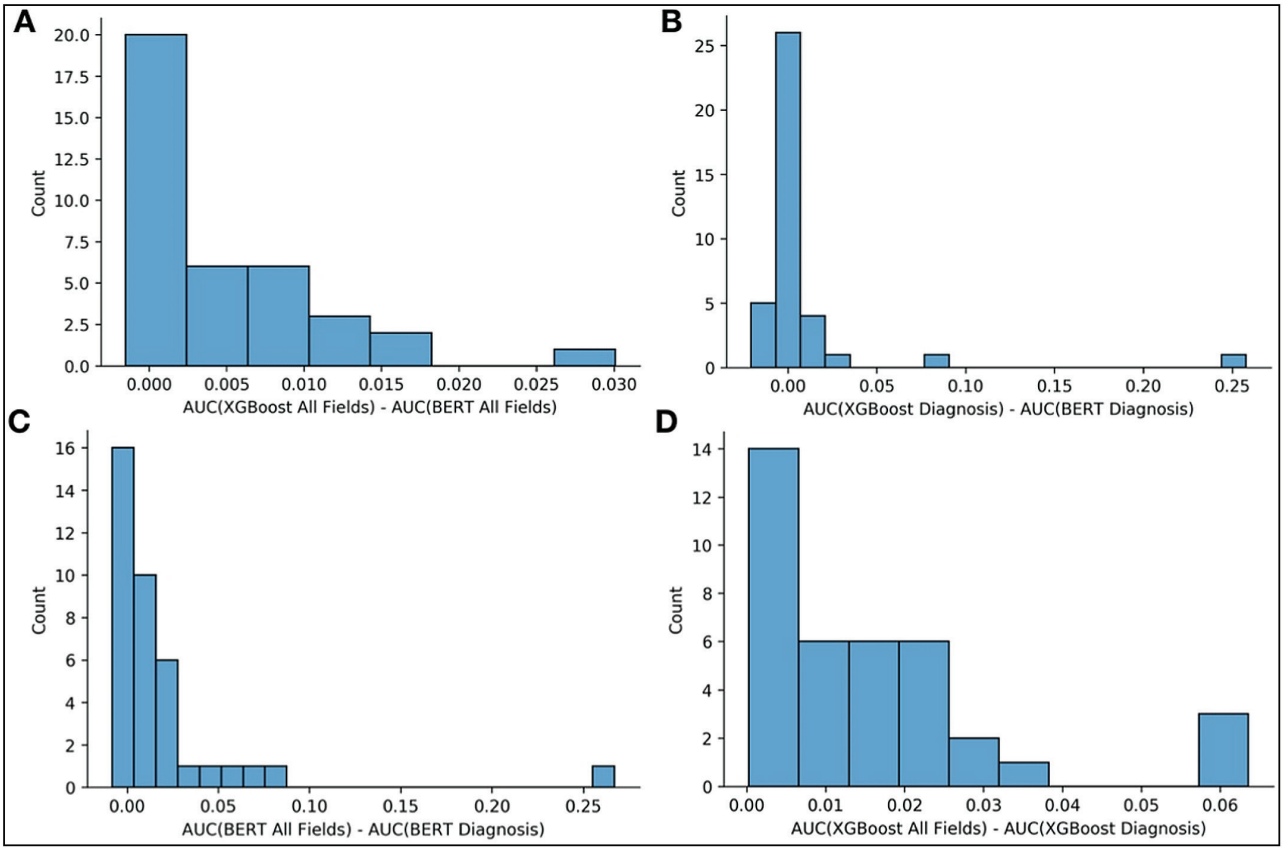
Supplementary Figure 6: Strength of correlation between CPT codes and HDBSCAN report clusters denoted by the size and color of each circle; large blue circles indicate strong positive associations, whereas large red circles indicate strong negative associations; associations for: **A)** diagnostic text; **B)** all-fields text



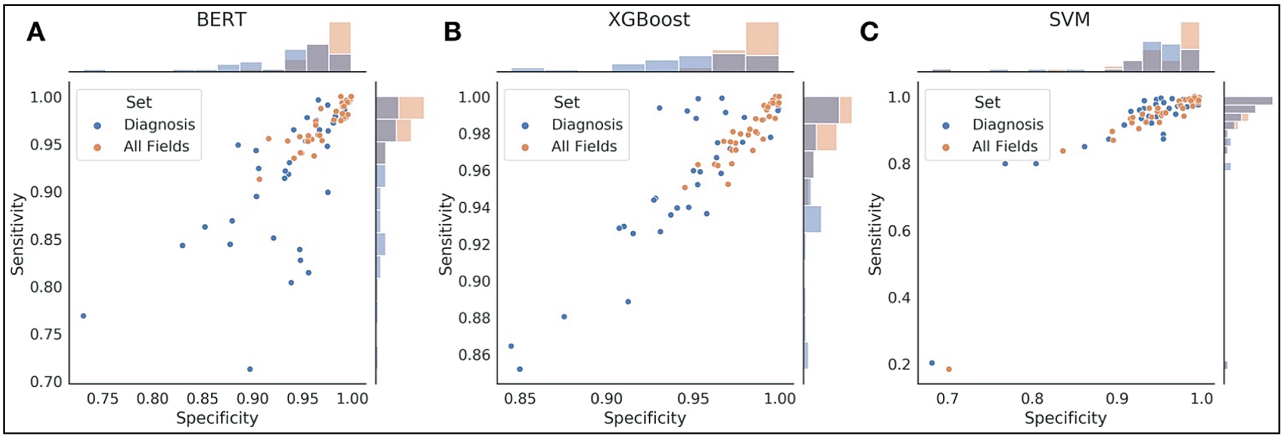
Supplementary Figure 7: Pathologist Associations: **A)** Clustered heatmap between associations/co-occurrence between pathologist and CPT codes establishes "subspecialties," where pathologists who order similar CPT codes are likely of similar subspecialty/subspecialties; left color track is colored by established subspecialty clusters; **B)** Clustered confusion matrix for pathologist prediction task (BERT diagnostic-text model); rows indicate true pathologists, whereas columns indicate predicted pathologist; row and column color bars utilize established "subspecialty" clusters; since the clustering of rows and columns place pathologists of a similar subspecialty together, this indicates that the misclassification occurred mostly within subspecialties



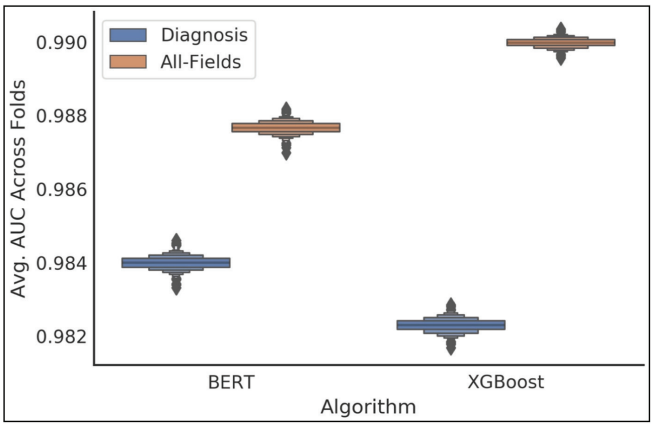
Supplementary Figure 8: Ancillary CPT Code Model Performance: Grouped boxenplots demonstrating the performance of machine-learning models (BERT, XGBoost, SVM) across CPT codes (distribution of AUCs reported for each CPT code), given the analysis of either the diagnostic text (blue) or all report subfields (orange)



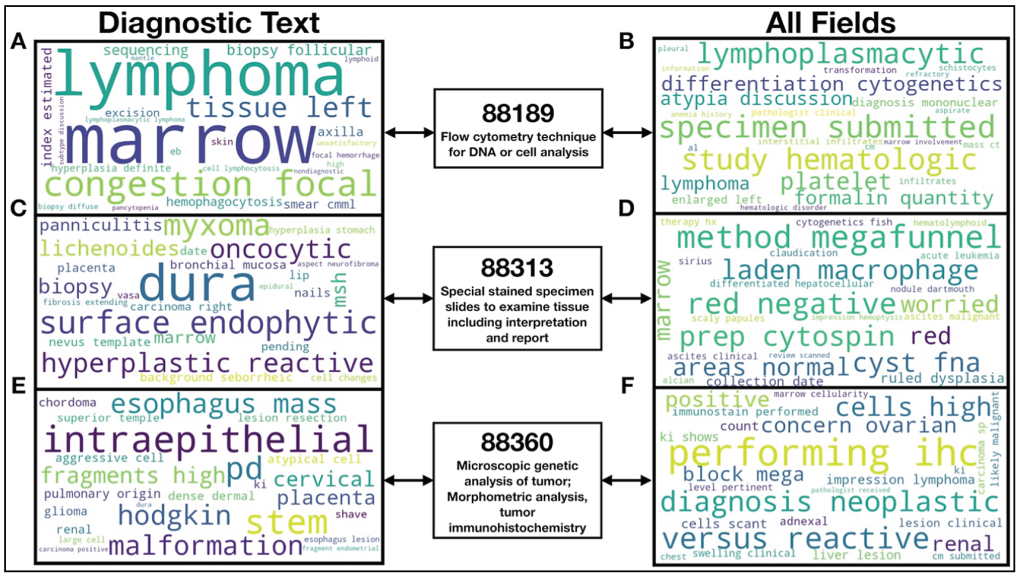
Supplementary Figure 9: Histogram of pairwise comparison (subtraction) of AUC statistics (averaged across cross-validation folds) between sets of algorithms / utilized document subfields; histogram tabulates AUC differences for individual codes, of which there are 38 values to be distributed among the histogram bins; reported relative performance gain (comparison/subtraction) of: **A)** XGBoost using all report subfields versus BERT using all report subfields, **B)** XGBoost using diagnostic subfield versus BERT using diagnostic subfield, **C)** BERT using all report subfields versus BERT using diagnostic subfield, **D)** XGBoost using all report subfields versus XGBoost using a diagnostic subfield



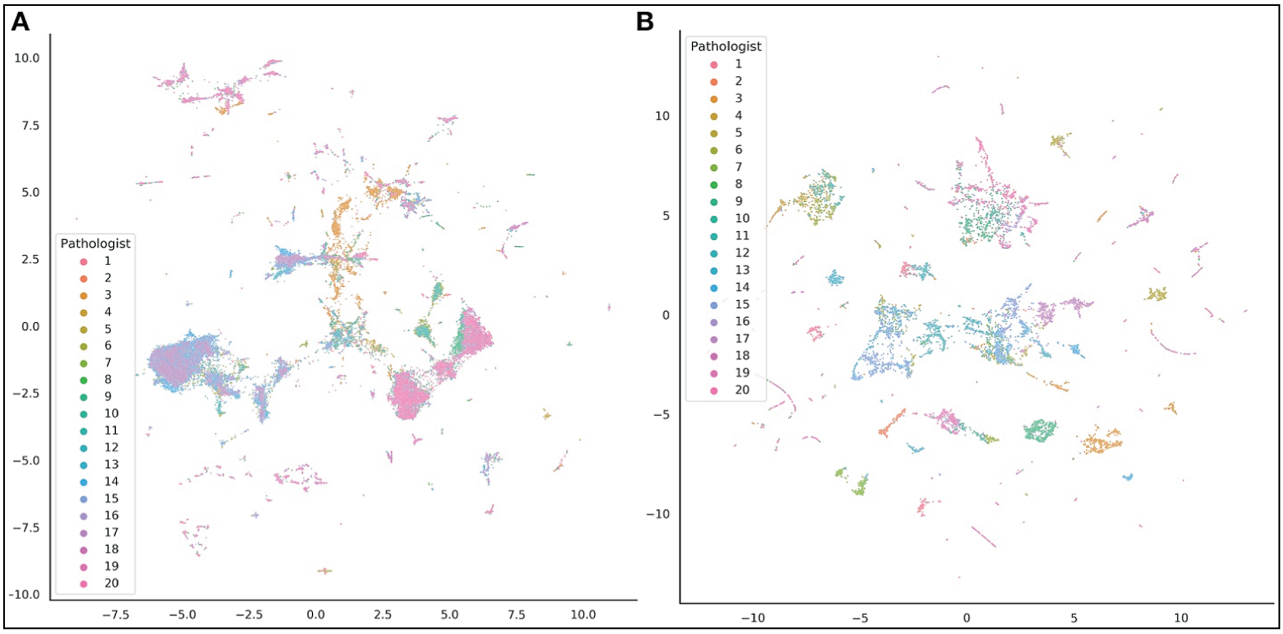
Supplementary Figure 10: Scatterplot of sensitivity and specificities for each CPT code, after averaging across CPT codes; the individual point is a CPT code; the point is colored by whether it was predicted from the diagnostic text or all report subfields; histograms at plot margins indicate marginal distribution of code sensitivity/specificity



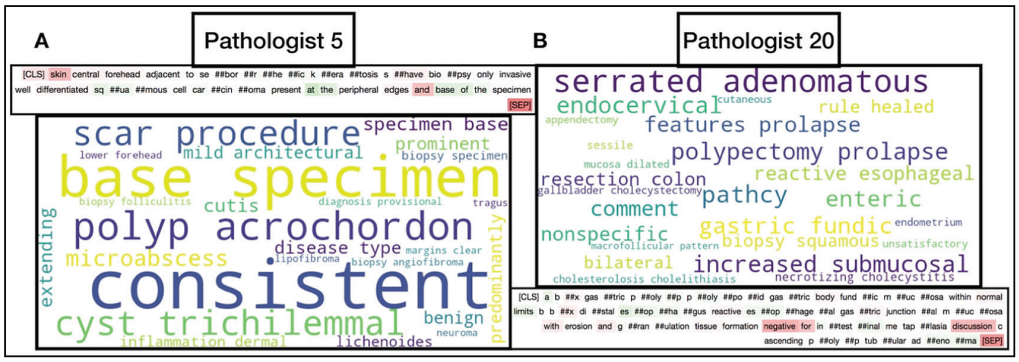
Supplementary Figure 11: Averaged weighted AUC statistics across pathologists/cross-validation folds for the prediction of top 20 pathologists with most sign-outs; reports for BERT and XGBoost for the diagnosis and all-fields models



Supplementary Figure 12: SHAP interpretation of XGBoost predictions: Word clouds demonstrating words found to be important using the XGBoost algorithm for the prediction of specific ancillary CPT codes, found via shapley attribution; important words pertinent to each CPT code indicated by the relative size of the word in the word cloud; word clouds visualized for three example CPT codes: **A-B)** CPT code 88189; **C-D)** CPT code 88313; **E-F)** CPT code 88360; visualizations performed for **A,C,E)** diagnostic text only, **B,D,F)** all report subfields (*all-fields*)



Supplementary Figure 13: Pathology reports colored by practicing pathologist: UMAP embeddings of pathology reports, colored by the pathologist who had written the report; each point indicates a pathology report, projected from use of either: **A)** Bag-Of-Words / tf-idf count matrix; **B)** embeddings after integrating information from all report subsections via the BERT *all-fields* model



Supplementary Figure 14: Interpretation of BERT and XGBoost models for pathologist prediction: Word cloud output of top words (size of word indicates importance; importance determined using SHAP) for XGBoost model prediction of the specific pathologist and Integrated Gradients highlighting of text via the BERT diagnostic model for select pathologists: **A)** Pathologist 5; **B)** Pathologist 20

Supplementary Table 1: Recording the percent missingness of each report subsection before removing reports lacking a diagnostic section. Summary measures (median, 1st quartile, 3rd quartile) for the number of words in each document subsection (where the subfield existed) and the percentage of documents whose length exceeded 512 words

Report subfield	Missingness before removal	Median word count	1st Q Word count	3rd Q Word count	Exceeds BERT max words
ADDENDUM DISCUSSION	96.2%	84	47	121	0.000%
ADDITIONAL STUDIES	86.6%	78	11	92	0.039%
CLINICAL INFORMATION	5.3%	25	18	61	0.000%
DIAGNOSIS	3.5%	23	13	28	0.022%
DISCUSSION	81.7%	36	16	68	0.017%
FINAL DIAGNOSIS	99.9%	235	186	313	0.000%
FROZEN SECTION	99.4%	2	1	11	0.000%
FROZEN SECTION DIAGNOSIS	99.3%	20	12	33	0.000%
INTERPRETATION	99.9%	135	91	141	0.000%
RESULTS	97.9%	248	216	268	2.198%
SPECIMEN PROCESSING	34.4%	38	27	64	0.389%
Complete Text (*All Fields*)	0%	119	68	158	1.768%

Supplementary Table 2: Changes in primary CPT code assignment over time; model fits for several logistic regression models, modeling time as years since 2017 (continuous) and whether a CPT code was assigned on a specific day as the dichotomous outcome variable

CPT Code	B	SE	P-value	CI [2.5%]	CI [97.5%]
88300	-0.050	0.037	0.172	-0.122	0.022
88302	0.135	0.053	0.012	0.030	0.239
88304	0.096	0.020	<0.001	0.056	0.136
88305	0.004	0.009	0.687	-0.014	0.021
88307	-0.004	0.018	0.818	-0.039	0.031
88309	-0.043	0.041	0.298	-0.123	0.038

Supplementary Table 3: Correlation between length of the word document and the number of uniquely assigned codes; broken down by a reported cluster using the diagnostic fields and all report fields

Cluster	Diagnostic clusters		All-field clusters	
	Correlation	p-value	Correlation	p-value
1	-0.09	1.6E-26	0.39	5.7E-178
2	0.07	1.6E-02	-0.05	7.1E-02
3	-0.30	3.4E-85	0.01	7.1E-01
4	-0.05	2.8E-02	0.02	3.9E-01
5	0.18	6.9E-93	0.00	9.6E-01
6	0.21	9.4E-37	0.01	2.9E-01
7	0.27	2.5E-26	0.08	1.7E-05
8	0.14	6.3E-137	0.57	4.9E-93
9			0.10	1.0E-28
10			0.31	5.3E-113
11			0.32	1.2E-33
12			0.16	1.0E-23
13			0.48	5.1E-98
14			0.09	3.9E-07
15			0.32	0.0E+00

Supplementary Table 4: Top 10 words found for each LDA topic ("topic descriptors"); 10 topics were discovered for the diagnostic text; and 10 additional topics were discovered for all of the report subfields (*All Fields*)

Diagnostic text	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
0	tumor	tissue	test	mucosa	cervical	colon	nevus	cells	shave	fragments
1	lymph	right	cancer	gastric	results	polypectomy	shave	placenta	cell	benign
2	carcinoma	left	lesion	esophagus	cancer	tubular	excision	cord	carcinoma	endocer-vical
3	grade	benign	cervical	chronic	please	adenoma	left	umbilical	left	evidence
4	nodes	excision	please	normal	guidelines	polyp	right	vessel	right	cervical
5	prostatic	soft	results	within	test	hyperplastic	melanocytic	acute	specimen	effect
6	left	breast	consensus	limits	consensus	ascending	changes	seen	basal	squamous
7	right	fallopian	management	abnormality	screening	sigmoid	compound	three	squamous	dysplasia
8	identified	inflam-mation	http://www.asccp.org	diagnostic	manage-ment	fragments	specimen	grams	discussion	mucosa
9	invasive	resection	guidelines	seen	cells	transverse	back	villous	peripheral	hpv
All-fields text	**Topic 1**	**Topic 2**	**Topic 3**	**Topic 4**	**Topic 5**	**Topic 6**	**Topic 7**	**Topic 8**	**Topic 9**	**Topic 10**
0	pap	skin	tissue	tissue	pap	tissue	biopsy	positive	clinical	skin
1	hpv	specimen	biopsy	polyp	hist	lymph	diagnosis	antibody	pertinent	shave
2	test	clinical	formalin	submitted	hpv	margin	specimen	tissue	total	biopsy
3	hist	'clock	quantity/size	colon	test	specimen	clinical	clinical	received	left
4	screening	excision	sections/processing	formalin	screening	tumor	see	studies	fluid	clinical
5	cervical	submitted	submitted	clinical	cervical	right	case	diagnostic	source	right
6	clinical	tissue	labeled/fixative	soft	clinical	left	punch	formalin	specimen	submitted
7	therapy	left	description	labeled/ fixative	therapy	node	discussion	staining	preparation	specimen
8	cancer	nevus	soft	history	cancer	submitted	submitted	core	description	tissue

Supplementary Table 5: Confusion matrices for each of the modeling approaches for primary CPT code prediction; aggregated across test sets of cross-validation folds; note how for BERT and XGBoost modes, misclassifications are mostly by codes of a similar case complexity

			Predicted									
			Diagnosis					All Fields				
			88302	88304	88305	88307	88309	88302	88304	88305	88307	88309
TRUE	**BERT**	88302	**357**	25	27	3	0	**356**	29	24	3	0
		88304	19	**3594**	322	92	2	18	**3568**	321	118	4
		88305	21	664	**50434**	392	20	16	614	**50490**	388	23
		88307	3	148	257	**3247**	36	3	131	253	**3249**	55
		88309	2	10	29	96	**123**	1	3	27	94	**135**
	XGBoost	88302	**338**	29	45	0	0	**334**	33	44	1	0
		88304	10	**3322**	616	78	3	12	**3418**	515	82	2
		88305	19	387	**50867**	250	8	8	376	**50921**	221	5
		88307	2	149	522	**2989**	29	0	116	366	**3178**	31
		88309	1	9	46	100	**104**	0	5	38	97	**120**
	SVM	88302	**45**	59	289	19	0	**56**	84	243	29	0
		88304	20	**3384**	436	189	0	3	**3312**	618	96	0
		88305	16	1516	**48712**	1287	0	13	1068	**49382**	1068	0
		88307	5	299	774	**2613**	0	11	373	736	**2571**	0
		88309	1	18	81	160	**0**	0	20	98	142	**0**

Supplementary Table 6: Summary of distribution of AUCs across ancillary CPT codes for BERT, XGBoost, and SVM prediction models for diagnostic and *all-fields* text

Model	Report subfields	Median	1st Quartile	3rd Quartile
BERT	Diagnosis	0.990	0.973	0.995
	All-Fields	0.995	0.985	0.999
XGBoost	Diagnosis	0.985	0.974	0.994
	All-Fields	0.997	0.994	0.999
SVM	Diagnosis	0.966	0.954	0.984
	All-Fields	0.977	0.957	0.992

Supplementary Table 7: Confidence intervals of 1000-sample nonparametric bootstrap of area under the receiver operating characteristic curve for each algorithm (BERT, XGBoost and SVM) and for each report type (Diagnosis and *All-Fields*); each AUC was averaged across the 5 cross-validation folds with the same random seed set for sampling values within each CV fold for each code/group of pathologists; ancillary CPT code and descriptions of codes listed on the left, in addition to the weighted AUC across 20 pathologists

		AUCs (± SE)					
		BERT		XGBoost		SVM	
Code	Description	Diagnosis	All-Fields	Diagnosis	All-Fields	Diagnosis	All-Fields
85060	Blood smear interpretation by physician with written report	0.998 ± 0.0002	0.9994 ± 0.0001	0.9989 ± 0.0002	0.9996 ± 0.0001	0.9983 ± 0.0002	0.9968 ± 0.0012
85097	Bone marrow, smear interpretation	0.9996 ± 0.0001	0.9994 ± 0.0001	0.9989 ± 0.0005	0.9997 ± 0.0	0.9985 ± 0.0001	0.9941 ± 0.0014
87491	Detection test for chlamydia	0.9905 ± 0.0008	0.9984 ± 0.0008	0.9898 ± 0.001	0.9996 ± 0.0002	0.9872 ± 0.0013	0.9819 ± 0.0042
87591	Detection test for Neisseria gonorrhoeae (gonorrhoeae bacteria)	0.9905 ± 0.0008	0.9994 ± 0.0001	0.9898 ± 0.001	0.9996 ± 0.0002	0.9872 ± 0.0013	0.9819 ± 0.0042
87624	Detection test for human papillomavirus (hpv)	0.9968 ± 0.0006	0.9973 ± 0.0003	0.9958 ± 0.0004	0.9984 ± 0.0002	0.9778 ± 0.0017	0.988 ± 0.0016
88108	Cell examination of specimen	0.9802 ± 0.0017	0.999 ± 0.0003	0.9808 ± 0.0008	0.9975 ± 0.0015	0.9717 ± 0.0026	0.9989 ± 0.0001
88112	Cell examination of specimen	0.9934 ± 0.0005	0.9991 ± 0.0001	0.9935 ± 0.0002	0.9995 ± 0.0	0.9887 ± 0.0004	0.9959 ± 0.0008
88141	Cytopathology, cervical or vaginal (any reporting system), requiring interpretation by physician	1.0 ± 0.0	0.9998 ± 0.0001	0.9996 ± 0.0001	0.9999 ± 0.0	0.9988 ± 0.0004	0.9923 ± 0.0014
88142	Pap test (Pap smear)	0.9886 ± 0.0017	0.9938 ± 0.0016	0.9826 ± 0.0017	0.9951 ± 0.0018	0.9663 ± 0.0018	0.9501 ± 0.0131
88172	Evaluation of fine needle aspirate	0.9825 ± 0.0011	0.999 ± 0.0002	0.9837 ± 0.0011	0.999 ± 0.0006	0.9749 ± 0.0015	0.9903 ± 0.001
88173	Evaluation of fine needle aspirate with interpretation and report	0.9867 ± 0.0024	0.9988 ± 0.0002	0.9899 ± 0.0005	0.9996 ± 0.0	0.9818 ± 0.001	0.997 ± 0.0006
88175	Pap test	0.998 ± 0.0005	0.9976 ± 0.0003	0.9972 ± 0.0003	0.9981 ± 0.0003	0.9932 ± 0.0009	0.9847 ± 0.002
88177	Pap test	0.9774 ± 0.0023	0.9993 ± 0.0001	0.9783 ± 0.0031	0.9998 ± 0.0	0.9624 ± 0.0044	0.9955 ± 0.0003
88184	Flow cytometry technique for DNA or cell analysis	0.9731 ± 0.0082	0.9848 ± 0.0022	0.9738 ± 0.0025	0.9942 ± 0.0012	0.9699 ± 0.0029	0.9708 ± 0.0033
88185	Flow cytometry, cell suXGBace, cytoplasmic, or nuclear marker, technical component only	0.9629 ± 0.0075	0.9841 ± 0.0022	0.9711 ± 0.0027	0.994 ± 0.0008	0.9594 ± 0.003	0.9692 ± 0.0034
88188	Cytopathology procedures	0.9428 ± 0.0121	0.9773 ± 0.0029	0.9589 ± 0.0041	0.9875 ± 0.0024	0.9593 ± 0.0041	0.9486 ± 0.0029
88189	Flow cytometry technique for DNA or cell analysis	0.9043 ± 0.0295	0.9753 ± 0.0052	0.9199 ± 0.0101	0.9785 ± 0.0073	0.9611 ± 0.0074	0.9471 ± 0.0118
88271	FISH DNA probe, each	0.9943 ± 0.002	0.9906 ± 0.0025	0.9735 ± 0.0055	0.995 ± 0.0024	0.9717 ± 0.0062	0.9768 ± 0.0061
88274	Genetic testing	0.9951 ± 0.0011	0.9943 ± 0.003	0.9755 ± 0.0059	0.9941 ± 0.0036	0.9775 ± 0.0058	0.9922 ± 0.0029
88300	Pathology examination of tissue using a microscope, limited examination	0.9983 ± 0.0011	0.9969 ± 0.0008	0.9967 ± 0.0012	0.9978 ± 0.0009	0.9846 ± 0.0025	0.9868 ± 0.0023
88302	Pathology examination of tissue using a microscope	0.9768 ± 0.0083	0.9824 ± 0.0036	0.9887 ± 0.0028	0.9934 ± 0.0019	0.9581 ± 0.0047	0.9643 ± 0.0042
88304	Pathology examination of tissue using a microscope, moderately low complexity	0.991 ± 0.0011	0.9877 ± 0.0007	0.987 ± 0.0009	0.9907 ± 0.0006	0.9534 ± 0.0019	0.9509 ± 0.0021

Supplementary Table 7: Contd

		AUCs (± SE)					
		BERT		XGBoost		SVM	
Code	Description	Diagnosis	All-Fields	Diagnosis	All-Fields	Diagnosis	All-Fields
88305	**Pathology examination of tissue using a microscope, intermediate complexity**	0.9726 ± 0.0012	0.9775 ± 0.0005	0.97 ± 0.0006	0.9889 ± 0.0003	0.1087 ± 0.0012	0.0807 ± 0.001
88307	**Pathology examination of tissue using a microscope, moderately high complexity**	0.9942 ± 0.0006	0.9928 ± 0.0004	0.9925 ± 0.0004	0.995 ± 0.0003	0.9614 ± 0.0015	0.968 ± 0.0013
88309	**Pathology examination of tissue using a microscope, high complexity**	0.9966 ± 0.0009	0.9885 ± 0.0021	0.9949 ± 0.0008	0.9967 ± 0.0007	0.9608 ± 0.0034	0.9777 ± 0.0022
88311	**Preparation of tissue for examination by removing any calcium present**	0.9906 ± 0.0033	0.9972 ± 0.0003	0.9943 ± 0.0009	0.9991 ± 0.0002	0.9316 ± 0.0035	0.9741 ± 0.0019
88312	**Special stained specimen slides to identify organisms, including interpretation and report**	0.9766 ± 0.0025	0.9792 ± 0.0012	0.9692 ± 0.0017	0.9972 ± 0.0004	0.8974 ± 0.0038	0.9063 ± 0.0031
88313	**Special stained specimen slides to examine tissue, including interpretation and report**	0.9577 ± 0.0065	0.9854 ± 0.0013	0.9619 ± 0.0023	0.9953 ± 0.0006	0.9163 ± 0.0039	0.9234 ± 0.0036
88321	**Surgical pathology consultation and report**	0.9945 ± 0.0007	0.998 ± 0.0007	0.9889 ± 0.001	0.9994 ± 0.0001	0.9483 ± 0.0033	0.9931 ± 0.0013
88331	**Pathology examination of tissue during surgery**	0.949 ± 0.0135	0.9958 ± 0.0012	0.9834 ± 0.0019	0.9996 ± 0.0002	0.9465 ± 0.0044	0.9592 ± 0.0024
88332	**Pathology examination of specimen during surgery**	0.8971 ± 0.0485	0.9821 ± 0.0063	0.974 ± 0.0059	0.9972 ± 0.0008	0.9077 ± 0.0186	0.9666 ± 0.0084
88333	**Pathology examination of tissue specimen during surgery**	0.9924 ± 0.0011	0.9963 ± 0.0018	0.9883 ± 0.0027	0.999 ± 0.0008	0.9827 ± 0.0021	0.979 ± 0.0076
88341	**Immunohistochemistry or immunocytochemistry, per specimen**	0.9353 ± 0.0034	0.96 ± 0.0012	0.9273 ± 0.0017	0.9901 ± 0.0004	0.8514 ± 0.0031	0.9262 ± 0.0022
88342	**Immunohistochemistry or immunocytochemistry, per specimen; initial single antibody stain procedure**	0.9384 ± 0.0024	0.9925 ± 0.0003	0.9319 ± 0.0011	0.9955 ± 0.0002	0.8404 ± 0.0021	0.9471 ± 0.0015
88344	**Special stained specimen slides to examine tissue**	0.9833 ± 0.0117	0.9824 ± 0.0075	0.9747 ± 0.0061	0.9942 ± 0.0028	0.9664 ± 0.0091	0.9627 ± 0.0091
88346	**Antibody evaluation**	0.9971 ± 0.0028	0.9972 ± 0.0018	0.9966 ± 0.0026	0.9977 ± 0.0023	0.987 ± 0.0045	0.989 ± 0.005
88350	**Antibody evaluation**	0.9999 ± 0.0001	0.9998 ± 0.0	0.9993 ± 0.0004	0.9999 ± 0.0	0.9852 ± 0.0048	0.9933 ± 0.0037
88360	**Microscopic genetic analysis of tumor; morphometric analysis, tumor immunohistochemistry**	0.7182 ± 0.0282	0.9853 ± 0.0022	0.9761 ± 0.0027	0.9944 ± 0.0013	0.9578 ± 0.0042	0.9564 ± 0.0048
	Top 20 pathologists	0.984 ± 0.0002	0.9877 ± 0.0002	0.9823 ± 0.0002	0.99 ± 0.0001	0.3778 ± 0.0007	0.3726 ± 0.0007

Supplementary Table 8: Wilcoxon tests for significance of relative performance gains (distribution of paired AUC differences for codes between two algorithms/report subfield combinations); all Wilcoxon tests were one-sided (algorithm 1 / selected subfields performance greater than algorithm 2 / selected subfields performance) to see which models perform the best for CPT code prediction

Algorithm 1		Algorithm 2		
Name	Report fields	Name	Report fields	P-Value
XGBoost	All fields	BERT	All fields	**2.8E-07**
XGBoost	Diagnosis	BERT	Diagnosis	6.4E-01
BERT	All fields	BERT	Diagnosis	**4.2E-05**
XGBoost	All fields	XGBoost	Diagnosis	**4.0E-08**
BERT	All fields	SVM	All fields	**4.0E-08**
BERT	Diagnosis	SVM	Diagnosis	**6.4E-05**
SVM	All fields	SVM	Diagnosis	**6.8E-03**

Supplementary Table 9: Sensitivity/specificity for each algorithm/report subfield(s), averaged across cross-validation folds for each CPT code after optimization of Youden's index to select the sensitivity/specificity

	BERT				XGBoost				SVM			
	Diagnosis		All fields		Diagnosis		All fields		Diagnosis		All fields	
Code	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity
85060	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
85097	0.98	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	1.00
87491	0.96	0.98	0.99	1.00	0.99	0.98	1.00	1.00	0.99	0.98	0.99	0.97
87591	0.99	0.98	0.99	1.00	0.99	0.98	1.00	1.00	0.99	0.98	0.99	0.97
87624	0.98	0.99	0.98	0.99	0.98	0.99	0.98	0.99	0.98	0.97	0.97	0.98
88108	0.84	0.95	0.99	0.99	0.99	0.95	0.99	1.00	0.99	0.95	1.00	0.99
88112	0.97	0.96	0.99	0.99	0.99	0.97	1.00	0.99	0.99	0.97	0.99	0.99
88141	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.99
88142	0.95	0.89	0.99	0.97	0.99	0.93	0.97	0.97	0.99	0.93	0.94	0.95
88172	0.81	0.96	0.99	0.99	1.00	0.95	0.99	0.99	1.00	0.95	0.99	0.98
88173	1.00	0.97	1.00	0.99	1.00	0.97	1.00	0.99	0.99	0.97	0.99	0.99
88175	0.98	0.99	0.98	0.99	0.98	0.99	0.98	0.99	0.98	1.00	0.97	0.98
88177	0.83	0.95	1.00	0.99	0.99	0.95	1.00	1.00	0.98	0.95	1.00	0.99
88184	0.85	0.92	0.95	0.96	0.94	0.93	0.98	0.98	0.96	0.93	0.96	0.94
88185	0.71	0.90	0.95	0.95	0.94	0.93	0.98	0.97	0.96	0.92	0.95	0.95
88188	0.87	0.88	0.95	0.94	0.93	0.91	0.96	0.96	0.96	0.93	0.92	0.93
88189	0.77	0.73	0.94	0.95	0.88	0.88	0.95	0.97	0.95	0.95	0.92	0.95
88271	0.92	0.94	0.95	0.97	0.94	0.95	0.98	0.99	0.96	0.96	0.97	0.98
88274	0.96	0.94	0.97	0.99	0.95	0.95	0.99	0.99	0.97	0.97	0.98	0.99
88300	0.99	0.99	0.98	1.00	0.99	0.99	0.99	0.99	0.97	0.98	0.96	0.98
88302	0.90	0.98	0.94	0.96	0.96	0.97	0.96	0.97	0.95	0.93	0.93	0.92
88304	0.95	0.96	0.96	0.96	0.96	0.95	0.96	0.96	0.93	0.93	0.92	0.92
88305	0.94	0.90	0.96	0.92	0.93	0.91	0.95	0.95	0.20	0.68	0.18	0.70
88307	0.97	0.97	0.97	0.96	0.97	0.96	0.98	0.97	0.94	0.94	0.95	0.93
88309	0.96	0.97	0.96	0.97	0.97	0.97	0.98	0.98	0.94	0.95	0.96	0.95
88311	0.97	0.98	0.98	0.98	0.98	0.98	0.99	0.99	0.89	0.95	0.97	0.95
88312	0.84	0.88	0.93	0.94	0.93	0.92	0.98	0.98	0.85	0.86	0.84	0.84
88313	0.89	0.90	0.94	0.95	0.89	0.91	0.97	0.97	0.87	0.89	0.87	0.90
88321	0.98	0.95	0.99	0.99	0.96	0.95	1.00	1.00	0.91	0.91	0.99	0.99
88331	0.93	0.94	0.98	0.98	0.94	0.96	1.00	1.00	0.92	0.93	0.94	0.92
88332	0.92	0.93	0.96	0.95	0.94	0.94	0.99	0.99	0.87	0.95	0.95	0.96
88333	0.95	0.95	0.98	0.99	0.97	0.96	1.00	1.00	0.98	0.96	0.97	0.99
88341	0.84	0.83	0.91	0.91	0.85	0.85	0.96	0.95	0.80	0.80	0.90	0.89
88342	0.86	0.85	0.97	0.96	0.86	0.84	0.98	0.97	0.80	0.77	0.90	0.93
88344	0.95	0.97	0.96	0.97	0.94	0.94	0.97	0.98	0.95	0.97	0.94	0.98
88346	0.92	0.91	0.99	0.99	0.99	1.00	1.00	1.00	0.98	0.97	0.98	1.00
88350	0.80	0.94	1.00	1.00	0.99	1.00	1.00	1.00	0.97	0.98	0.99	1.00
88360	0.91	0.93	0.95	0.96	0.93	0.93	0.97	0.97	0.92	0.94	0.94	0.94

Supplementary Table 10: Additional performance statistics: First three numerical columns: Averaged sensitivity and specificity across the XGBoost and BERT algorithms to denote overall predictive performance for each CPT code; Average Youden calculated from the sensitivity and specificity; Final three numerical columns: Changes in sensitivity, specificity, and Youden when utilizing all report subfields versus the diagnostic text alone

Code	Average sensitivity	Average specificity	Average Youden	Δ Sensitivity	Δ Specificity	Δ Youden
85060	1.00	1.00	0.99	0.00	0.00	0.00
85097	0.99	0.99	0.99	0.01	0.01	0.02
87491	0.99	0.99	0.97	0.02	0.02	0.04
87591	0.99	0.99	0.98	0.00	0.02	0.02
87624	0.98	0.99	0.97	0.00	0.00	0.00
88108	0.95	0.97	0.92	0.08	0.04	0.12
88112	0.99	0.98	0.97	0.01	0.02	0.04
88141	1.00	1.00	1.00	0.00	0.00	0.00
88142	0.98	0.94	0.92	0.01	0.06	0.07
88172	0.95	0.97	0.92	0.08	0.04	0.12
88173	1.00	0.98	0.98	0.00	0.03	0.03
88175	0.98	0.99	0.97	0.00	0.00	0.00
88177	0.95	0.97	0.93	0.09	0.05	0.14
88184	0.93	0.95	0.88	0.07	0.04	0.11
88185	0.90	0.94	0.84	0.14	0.05	0.19
88188	0.93	0.92	0.85	0.06	0.05	0.11
88189	0.89	0.88	0.77	0.12	0.16	0.28
88271	0.95	0.96	0.91	0.04	0.04	0.07
88274	0.97	0.97	0.94	0.02	0.04	0.06
88300	0.99	0.99	0.98	0.00	0.00	0.00
88302	0.94	0.97	0.91	0.02	0.00	0.02
88304	0.96	0.96	0.92	0.00	0.00	0.01
88305	0.94	0.92	0.86	0.02	0.03	0.04
88307	0.97	0.97	0.94	0.01	0.00	0.01
88309	0.97	0.97	0.94	0.00	0.00	0.01
88311	0.98	0.98	0.97	0.02	0.01	0.02
88312	0.92	0.93	0.85	0.07	0.07	0.14
88313	0.92	0.93	0.86	0.06	0.05	0.12
88321	0.98	0.97	0.96	0.03	0.04	0.07
88331	0.96	0.97	0.93	0.05	0.04	0.09
88332	0.95	0.95	0.90	0.04	0.03	0.07
88333	0.98	0.98	0.95	0.03	0.03	0.06
88341	0.89	0.88	0.78	0.09	0.09	0.18
88342	0.92	0.91	0.83	0.11	0.12	0.23
88344	0.95	0.97	0.92	0.02	0.02	0.04
88346	0.98	0.97	0.95	0.03	0.04	0.08
88350	0.95	0.98	0.93	0.10	0.03	0.13
88360	0.94	0.95	0.89	0.04	0.03	0.08

Supplementary Table 11: Classification reports for pathologist prediction models (BERT, XGBoost, SVM) for reported subfields (diagnostic/all fields)

BERT							
Diagnosis				All fields			
Pathologist	Precision	Recall	F1-Score	Pathologist	Precision	Recall	F1-Score
1	0.94	0.94	0.94	1	0.95	0.94	0.94
2	0.49	0.82	0.61	2	0.61	0.84	0.70
3	0.94	0.86	0.89	3	0.99	0.98	0.98
4	0.77	0.76	0.77	4	0.81	0.81	0.81
5	0.80	0.85	0.82	5	0.88	0.88	0.88
6	0.93	0.98	0.95	6	0.96	0.96	0.96
7	0.81	0.82	0.81	7	0.87	0.87	0.87
8	0.36	0.91	0.51	8	0.41	0.80	0.55
9	0.86	0.78	0.82	9	0.86	0.80	0.83
10	0.78	0.61	0.69	10	0.74	0.68	0.71
11	0.67	0.71	0.69	11	0.71	0.73	0.72
12	0.84	0.77	0.80	12	0.87	0.83	0.85
13	0.80	0.91	0.85	13	0.86	0.91	0.88
14	0.72	0.74	0.73	14	0.83	0.85	0.84
15	0.83	0.74	0.78	15	0.84	0.83	0.83
16	0.56	0.25	0.34	16	0.54	0.35	0.42
17	0.89	0.96	0.93	17	0.93	0.96	0.94
18	0.58	0.14	0.22	18	0.45	0.27	0.34
19	0.71	0.72	0.71	19	0.71	0.74	0.72
20	0.84	0.39	0.53	20	0.74	0.43	0.54
Accuracy	0.74	0.74	0.74	Accuracy	0.79	0.79	0.79
Macro Avg	0.76	0.73	0.72	Macro Avg	0.78	0.77	0.77
Weighted Avg	0.77	0.74	0.74	Weighted Avg	0.80	0.79	0.79
XGBoost							
Diagnosis				All fields			
Pathologist	Precision	Recall	F1-Score	Pathologist	Precision	Recall	F1-Score
1	0.92	0.88	0.90	1	0.94	0.89	0.91
2	0.67	0.66	0.67	2	0.68	0.76	0.72
3	0.90	0.85	0.88	3	1.00	1.00	1.00
4	0.81	0.76	0.78	4	0.80	0.83	0.81
5	0.74	0.89	0.81	5	0.86	0.91	0.88
6	0.94	0.98	0.96	6	0.97	0.98	0.97
7	0.88	0.77	0.82	7	0.92	0.88	0.90
8	0.36	0.87	0.51	8	0.51	0.73	0.60
9	0.72	0.88	0.79	9	0.79	0.86	0.82
10	0.80	0.62	0.70	10	0.76	0.67	0.72
11	0.75	0.77	0.76	11	0.78	0.76	0.77
12	0.73	0.81	0.77	12	0.79	0.87	0.83
13	0.83	0.76	0.79	13	0.92	0.87	0.90
14	0.78	0.68	0.73	14	0.91	0.83	0.87
15	0.75	0.73	0.74	15	0.83	0.82	0.82
16	0.50	0.32	0.39	16	0.56	0.47	0.51
17	0.69	0.52	0.59	17	0.88	0.76	0.81
18	0.69	0.21	0.32	18	0.58	0.42	0.48
19	0.71	0.74	0.72	19	0.71	0.75	0.73
20	0.83	0.38	0.53	20	0.70	0.51	0.59
Accuracy	0.73	0.73	0.73	Accuracy	0.80	0.80	0.80
Macro Avg	0.75	0.70	0.71	Macro Avg	0.79	0.78	0.78
Weighted Avg	0.75	0.73	0.72	Weighted Avg	0.80	0.80	0.80

Supplementary Table 11: Contd....

BERT							
Diagnosis				All fields			
Pathologist	**Precision**	**Recall**	**F1-Score**	**Pathologist**	**Precision**	**Recall**	**F1-Score**
SVM							
Diagnosis				All fields			
Pathologist	Precision	Recall	F1-Score	Pathologist	Precision	Recall	F1-Score
1	0.59	0.62	0.60	1	0.45	0.50	0.47
2	0.38	0.36	0.37	2	0.10	0.00	0.00
3	0.56	0.57	0.56	3	0.86	0.84	0.85
4	0.33	0.16	0.22	4	0.20	0.18	0.19
5	0.39	0.52	0.44	5	0.24	0.73	0.36
6	0.36	0.56	0.44	6	0.34	0.65	0.45
7	0.09	0.04	0.05	7	0.00	0.00	0.00
8	0.36	0.80	0.49	8	0.34	0.92	0.49
9	0.49	0.67	0.57	9	0.28	0.79	0.41
10	0.34	0.09	0.14	10	0.18	0.05	0.07
11	0.44	0.32	0.38	11	0.23	0.32	0.26
12	0.24	0.36	0.29	12	0.21	0.24	0.23
13	0.00	0.00	0.00	13	0.00	0.00	0.00
14	0.00	0.00	0.00	14	0.00	0.00	0.00
15	0.23	0.42	0.30	15	0.26	0.11	0.16
16	0.30	0.18	0.23	16	0.28	0.04	0.07
17	0.00	0.00	0.00	17	0.00	0.00	0.00
18	0.06	0.02	0.02	18	0.18	0.03	0.05
19	0.32	0.49	0.38	19	0.00	0.00	0.00
20	0.07	0.02	0.03	20	0.11	0.00	0.00
Accuracy	0.35	0.35	0.35	Accuracy	0.32	0.32	0.32
Macro Avg	0.28	0.31	0.28	Macro Avg	0.21	0.27	0.20
Weighted Avg	0.29	0.35	0.30	Weighted Avg	0.24	0.32	0.24

Supplementary Table 12: SHAP coefficients depicting relationships between the top 30 words that distinguish the primary CPT codes and their related CPT code: Positive value indicates positive association, whereas negative value indicates negative association between word and code; top codes determined by summing absolute SHAP value across CPT codes and test cohort

	88302	88304	88305	88307	88309
Myocyte				2.337	
Excision pilomatricoma		-1.537	0.006		
Endocervical		-1.515		0.001	0.0
Ureter fresh				1.313	
Left ankle		0.45	-0.813		
Products conception		0.029		-1.161	
Biopsy	-0.376	0.054	0.466	0.235	0.045
Specimen cm		-1.059	0.108		
Mesh	0.201	-0.001	-0.929		
Spleen			0.0	-1.085	
Diagnosis skin		0.025	0.006	0.168	-0.836
Reduction			1.081		
Termination		0.234	-0.846		
Toe clinical				-1.044	
Mucocele		1.013	-0.001		
Hemorrhoid		0.818	-0.177		
Fixative pilonidal		0.958	0.0		
Valve		-0.945	0.004		
Irregular	0.217	0.684	-0.048	-0.003	
Representative	0.032	0.259	0.484	0.015	0.148
Metatarsal resection			0.937		
Submitted skin	-0.69	0.064	-0.044	-0.118	
Angioleiomyoma		-0.358		0.54	
Ovary serous				0.897	
Foreskin clinical		0.879			
Capsule excision	0.878				
Dcagnosis fibroma				0.874	
Transected	0.658		0.046	-0.159	
Mass provided			-0.756	0.083	
Excision suggestive		-0.819			